International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

# Data-Driven Clinical Decision Support System for Medical Diagnosis and Treatment Recommendation

**Shubham Rathi, Mahesh Motwani, Manish Ahirwar**

***Abstract*: This paper presents a Data-Driven Clinical Decision Support System (CDSS) using machine learning. The proposed system predicts the possibility of diseases based on the patient's symptoms. It suggests lab tests and medication related to the disease. Lab test results are analyzed to check the probability of liver and kidney diseases. The proposed system uses face recognition to identify the patient. Face recognition module retrieves the Patient Health Record and provides patient information and health records access to the doctor and medical staff. The system is developed using Python Django for Backend, React.JS for User Interface and PostgreSQL as the relational database. The system uses Logistic Regression for possible disease prediction, Support Vector Machine for liver disease prediction, Random Forest for chronic kidney disease prediction. The result of the proposed data-driven clinical decision support system is compared with a doctor's disease analysis to measure the effectiveness of the proposed system. This kind of system can help doctors in providing better care and predict the disease at an early stage.***

***Keywords*: Chronic Kidney Disease Prediction, Clinical Decision Support System, Disease Prediction, Machine Learning, Liver Disease Prediction**

## I. INTRODUCTION

M ost hospitals and doctors have transitioned from paper-based medical records to electronic health records (EHR)[1] that improves the efficiency of the doctor by documenting the treatment and laboratory results. The use of EHR gives medical professionals and patients complete information about past treatment and medication history.

Patient identification is the first and important step to proceed further in the hospital ecosystem. The patient identification process is usually done by identifying the patient by mobile number or name, and age. Accessing the correct record in the hospital system depends on the correct identification of the Patient. Failure to identify the patient may harm the patient health by providing incorrect medication based on the previous records, wrong patient medical record updates and money loss to both patient and hospital.

**Manuscript published on 30 September 2019.**
*****Correspondence Author(s)
**Shubham Rathi**, Computer Science and Engineering, University Institute of Technology, RGPV, Bhopal (M.P), India.
**Mahesh Motwani**, Computer Science and Engineering, University Institute of Technology, RGPV, Bhopal (M.P), India.
**Manish Ahirwar**, Computer Science and Engineering, University Institute of Technology, RGPV, Bhopal (M.P), India.



In the proposed system, Face recognition for patient identification is used to improve the efficiency of accessing medical records and to reduce patient misidentification. The use of face recognition for patient identification will help the hospital in retrieving the patient information and medical records in case of an emergency where the patient may be in an unconscious state.

While the Healthcare system is digitizing, Clinical decisions are taken on the basis of a doctor's knowledge and experience. A large number of doctors acquires medical information from different sources and then aided by experience makes the decision of possible disease. The whole process takes place in the brain of the doctor. In some cases, doctors might not be able to diagnose it accurately which may lead to wrong treatment.

To prevent this, it is necessary to utilize the amount of medical data available and help clinicians in decision making by suggesting the possible diseases. Machines can assist doctors by presenting patient-related specific information for better clinical decisions. The clinical decision support system uses the medical knowledge from different sources and assists the doctor in diagnosis, medication, treatment of the patient. The purpose of a CDSS[2] is to assist healthcare providers, by analyzing patient data and apply the medical knowledge on patient data to improve the accuracy of diagnosis.

The proposed data-driven clinical decision support system takes the patient's symptoms as input from the doctor and uses the logistic regression algorithm to predict the possible diseases. The system shows the 5 diseases that have the highest probability on the basis of given symptoms. Along with disease, It also recommends the lab test and radiology tests that need to perform to confirm the suggested disease. The system suggests the medication of the predicted disease from a medicine dataset[3] to help the doctor in prescribing the correct medicine. To analysis the lab test results, a Support Vector Machine algorithm is used to predict the probability of liver disease and the Random Forest algorithm is used to predict the probability of kidney disease.

## II. LITERATURE REVIEW

The variety of clinical decision support systems are available varies from personal digital assistant applications developed for a single hospital to mainframe-based systems. In addition, many researchers have developed CDSS according to research aims. Some of them CDSS available and used by doctors are discussed below:

**Clinical Rules**[4] is a real-time decision support module developed by Digitalis Rx. It focuses on medication safety and medicine optimization. Clinical Rules use a Knowledge-based Clinical Rules Engine for medication analysis. It uses patient data, lab test reports, prescribed medication, age, and gender for analysis and provides personalized advice to the doctor. Fig. 1 shows the process of clinical rules. Clinical Rules System divided into two parts: Clinical Rule Reporter (CRR) and Clinical Rule Engine (CRE). Clinical Rule Reporter sends the patient data to the Clinical Rule Engine from Electronic Health Records. Clinical Rule Engine performs the analysis on patient data and sends the report and alert information back to Clinical Rule Reporter. Clinical Rules are used for medication-related alert and analysis.

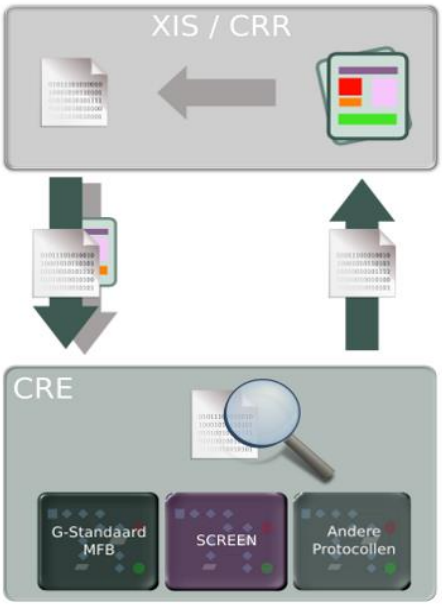


**Fig. 1. Process of Clinical Rule[4]**

**Infermedica**[5] uses Artificial Intelligence and Machine Learning to predict the disease on the basis of symptoms and lab test results. It utilizes the previous medical records to train a machine learning algorithm. Its AI Engine has an accuracy of 93%. It provides the REST APIs to healthcare organizations to utilize its platform and integrate it into the hospital system and mobile applications. It has a dataset of more than 1500 symptoms and 800 medical conditions, and support natural language processing in more than 15 languages.

**ESAGIL**[6] is an online free web application that matches symptoms, blood test values, and urine test values against the disease dataset. It has a dataset of more than 100 diseases and disorders. It is a Knowledge-based Decision Support System. It doesn't provide any application programming interface to integrate it into the hospital system. It can be used through the website only. The symptoms need to select from a list of more than 150 symptoms and after that lab test values need to choose between normal and abnormal according to the reference value. It returns the list of all the possible diseases and matching percentage of disease.

**ISABEL**[7] is 25 years old US Based company that has 4 products: Isabel Pro, Symptoms Checker, Active Intelligence and Clinical Educator. Isabel pro is a Differential Diagnosis (DDx) Generator that provides support for more than 10,000 diagnoses. Isabel pro takes the patient detail, symptoms and lab test reports and returns the list of possible diseases and list of drugs that can cause the problem. It also suggests the lab test and treatment to the doctor. Isabel is an evidence-based knowledge-based CDSS that provided REST API and SOAP API in HL7 Standard. It is a paid service. It provides Software as a Service. Active Intelligence system uses Artificial Intelligence and Natural Language Processing to convert the text of electronic health record (EHR) and electronic medical record (EMR) into a structured format to perform diagnosis.

**SimulConsult**[8] has support for more than 7,000 diagnoses, mostly focus on genetics, neurology and pediatric related diseases. It is knowledge-based CDSS which also uses an AI-based statistical pattern matching approach for analyzing by age of onset and offset finding. It provides a portal to access SimulConsult. Fig. 2 shows the User Interface of SimulConsult Web Application. SimulConsult returns the list of diseases sorted by probability. To improve the probability, it suggests the tips and lab tests in the 'Add Findings' tab. An advanced version of SimulConsult performs the disease diagnosis on Genes.

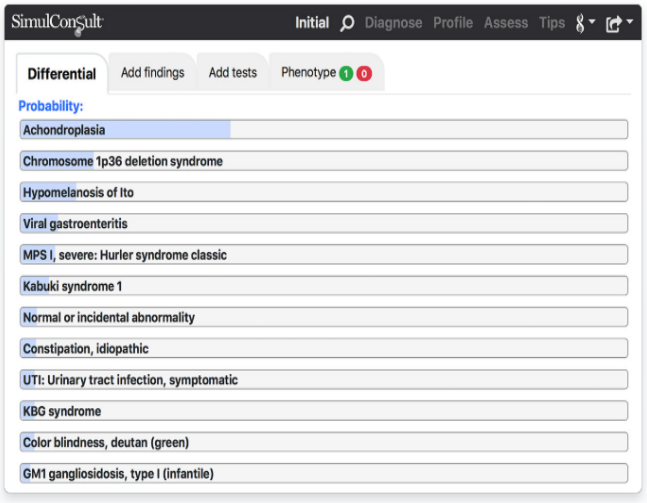


**Fig. 2. User Interface of SimulConsult[8]**

Table I shows a comparison of various clinical decision support systems available online. The systems are compared on the basis of their types, application interface type and their usage (field).

## III. METHODOLOGY

### A. Proposed Solution

In the hospital system, the first step is to verify the patient by his/her identity and search for the previous medical record. Medical staff verify the patient identity and book the consultant. After that, the doctor may search for the patient's medical record in the hospital system. At each and every step, Patient Identification and medical record retrieval process are repeated. To optimize the time consumption in record retrieval and reduce the case of patient misidentification, the face recognition system is used for patient identification.

Along with Face Recognition, Clinical Decision Support System is proposed to assist the doctor in finding all the possibility of diseases based on the patient's symptoms. CDSS also

**Table-I : Comparison Table of Various Clinical Decision Support System**

| S.No | CDSS | CDSS Type | Application Type | Application Field |
|---|---|---|---|---|
| 1 | **Clinical Rules[4]** | Knowledge-based | Web Application, Web Agents | Medicine prescription and consumption monitoring |
| 2 | **Infermedica[5]** | Machine Learning and Artificial Intelligence | Rest API | Diagnosis of more than 800 medical conditions. |
| 3 | **ESAGIL[6]** | Knowledge-based | Web Application | Diagnosis of 100 diseases according to symptoms, blood, and urine test results. |
| 4 | **ISABEL[7]** | Knowledge-based | Web Apps, Mobile App, REST API | Diagnosis of 10000 diseases |
| 5 | **SimulConsult[8]** | Knowledge-based, bioinformatics genome annotation, Artificial Intelligence based statistical pattern- matching approach | Web App | Diagnosis of 7184 diseases especially genetic and neurological diseases. |

performance the lab test analysis on the patient's lab test results to predict the probability of liver and kidney disease.

Fig. 3 shows the process flow diagram of the proposed system. When a patient enters the hospital or meets the doctor, the first step is to identify the patient using face recognition. After that, the Doctor or Medical Staff can select the patient's symptoms from the list of symptoms in the system. The selected symptoms are passed to the developed data-driven clinical decision support system to predict the possible diseases. A list of possible diseases with their probability based on symptoms is returned by the system.

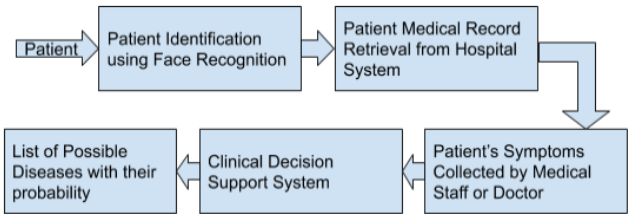


**Fig. 3 Process flow of Proposed Solution**

The proposed system uses multiple datasets and different machine learning algorithms to perform predictions on the patient's symptoms and lab test results. The dataset [9] from kaggle.com [10] is used to train the Logistic Regression algorithm to predict the probability of diseases based on symptoms provided by the doctor. Liver Disease dataset[11] from UCI[12] is used to train the Support Vector Machine algorithm to predict the probability of liver disease and Chronic Kidney Disease dataset[13] from UCI[12] is used to train Random Forest algorithm to predict the probability of kidney diseases by analyzing the lab test reports. A medicine dataset[3] from UCI[12] is used to suggest medicines according to disease. A deep learning library dlib[14] and python library face_recogntion[15] is used to implement face recognition for patient identification. Detailed implementation steps are discussed in the next sections.

### B. Data-Driven Clinical Decision Support System

Fig. 4 shows the architecture of the Data Driven Clinical Decision Support System. The system takes the list of symptoms and patient lab test results as an input and perform prediction on input data and return the list of possible disease with their probability. The system is split into 2 modules.

1) Early Stage Disease Prediction
2) Lab Test Result Analysis

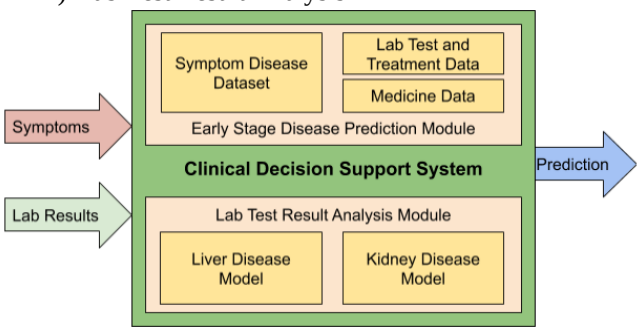


**Fig. 4 Architecture of Data-Driven CDSS**

#### 1) Early Stage Disease Prediction

An early-stage disease prediction system is developed to suggest possible diseases based on the patient's symptoms. It shows the list of possible diseases to doctor with the probability of each disease. It can help in providing a different angle of diagnosis which doctor might not have thought of in a busy schedule.

**Dataset and Machine Learning Algorithms Used:**

**Symptom Disease Dataset:** A dataset[9] from Kaggle[10] containing 4920 rows is used. Dataset has 41 unique diseases and 132 symptoms. We performed the preprocessing and data cleaning on the dataset before using the dataset for machine learning training.

We removed all the rows which contain the null value or empty value. After data cleaning, we split the dataset into a training dataset and test dataset. Multinomial Naive Bayes (MNB), Decision Tree (DT), and Logistic Regression (LR) algorithms are trained using a training dataset and performance of algorithms are compared. The trained Logistic Regression model is saved in a pickle file to predict the disease based on the patient's symptoms.

**Medicine Recommendation Dataset:** A dataset[3] from UCI repository[12] containing 161297 rows is used to prepare the medicine recommendation data set. The columns of the dataset are *drug name, condition, review, rating, usefulCount, and date.* We perform preprocessing and data cleaning on the dataset. Dropped all the rows which contain a null value, special symbols or empty value. We deleted the rows which have a '*rating'* smaller than 5 and '*useful count'* is smaller than 10. We were not using the values of review and date column, so we dropped the review and date column. An average score is calculated for each medicine by performing a product of '*rating'* and '*usefulCount'*. The medicines are grouped according to disease name *'condition'* column and sorted by average score. The higher the average score, the better the medicine's effectiveness. The processed Medicine Recommendation Dataset contains the 5330 rows which have 2428 unique medicines and 614 unique disease names.

**Lab Test and Treatment Recommendation Dataset:** A dataset of lab tests and treatment of 41 diseases is handcrafted by using the information from WebMD[16], Cleveland Clinic[17], UCSF Health[18]. The dataset has following fields: list of symptoms, disease name, test stage 1 (List of Primary Suggested Lab Tests), test stage 2 (List of Additional Lab Test), and treatment-related suggestion for few diseases.

**Algorithm for Early Stage Disease Prediction and Medication, Treatment and Lab Test Recommendation on the basis of Patient's Symptoms:**

**INPUT:** List of Symptoms

**OUTPUT:** List of diseases with its treatment, medicines and lab tests recommendation.

**Steps of Algorithm:**

1. Dataset and Machine Learning Model Loading

All the above-discussed datasets and trained logistic regression model from the pickle file are loaded in memory. This is the initialization setup of the Algorithm.

2. Input Preprocessing

The system generates an array of size 132 in which a symptom marked as 1 if the symptom is in the list of input symptoms else it is marked 0.

3. Disease Prediction

The generated array passed to the machine learning model to predict the probability of the disease. The machine learning model returns a list of 41 diseases with the probability of each disease on the basis of given symptoms.

4. Select Top 5 Diseases

A list of diseases is sorted according to the decreasing order of probability of disease. The top 5 diseases are selected. Step 5 and Step 6 are repeated for each disease in selected diseases.

5. Treatment and Lab Test selection from Dataset

The recommended treatment and lab tests are selected from the prepared lab test and treatment recommendation dataset for each disease.

6. Medicine Recommendation from Dataset

The predicted disease name is searched in the medicine dataset. If a disease found in a dataset then a maximum of 5 medicines are suggested for each disease.

7. Result

Return the list of diseases with probability, treatment, lab test, and medicine recommendation is returned to the user.

**2) Lab Test Result Analysis**

Lab test result analysis required a large set of rules for every lab test condition. To perform, lab test analysis, We have developed liver disease and kidney disease classification model by using the Chronic Kidney Disease Dataset[13] and Liver Disease Dataset[11] of UCI[12] repository. The patient lab reports data is used to perform kidney disease prediction and liver disease prediction. The lab test result analysis uses two machine learning model:

a) Kidney Disease Prediction

b) Liver Disease Prediction

**a) Kidney Disease Prediction:**

The Kidney Disease datasets set contains data of 400 patients in which 250 patients are marked 'ckd' which means the patient is suffering from chronic kidney disease and 150 patients are marked 'nckd'. The dataset has 25 columns in which 24 columns are values for classification and one column defines the classification class i.e 'ckd' or 'nckd'.

Following Process is used to analyze and select the machine learning model:

1. Data preprocessing and cleaning: We removed all the rows which contain a null value. Few rows contain 'ckd/d' instead of 'ckd' in the classification column. Replaced all such fields with 'ckd'. The processed dataset reduced to 158 rows after all data cleaning and preprocessing.
2. Data normalization: In this step, we replace the 'ckd' with 1 and 'nckd' with 0 in the classification field. Similarly, all the 14 columns containing nominal values are converted into a numerical value in the form of 0 or 1.
3. Data splitting: Dataset is split into training and test data. The size of the training dataset is 110 rows and the test data set is 48 rows.
4. Machine model train and measurement: Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), K-nearest neighbor (KNN) algorithms are trained using training dataset and accuracy score, precision, recall, fscore of each algorithm is measured on the test dataset.
5. Machine Learning Model Analysis: The results and performance comparison of the algorithms are discussed in the results section of the paper. After analysis, we found that the Random Forest algorithm performs better for chronic kidney disease prediction. Trained random forest model is saved in pickle file to perform prediction in the Django REST server on a patient's lab test value.

**b) Liver Disease Prediction:**

Liver disease dataset contains 583 rows and 11 columns in which 416 rows are of the liver patient and 167 rows are of the non-liver patient.

Following process is used to compare and select the machine learning algorithm:

1. Data cleaning: Removed all the rows containing null values. The processed dataset contains 579 rows after all null rows removal.
2. Data normalization: The gender column is replaced with two new columns “male”, “female”. The patient whose gender value is “male” is marked 1 in the male column and 0 in the female column and vice versa for the female gender.
3. Data Splitting: Dataset is split into x_train, x_test. The size of the training (x_train) dataset is 405 rows and the test data set is 174 rows.
4. Machine model train and measurement: Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM), K-nearest Neighbor (KNN) algorithms are trained using a training dataset. To compare the best algorithm for classification, Algorithm is compared using accuracy score, precision, recall, fscore on the test dataset.
5. Machine Learning Model Analysis: Performance comparison of the algorithm is discussed in the results section. We found that the Support Vector Machine algorithm performs better as compared to other machine learning algorithms for liver disease prediction. The trained SVM model is saved in the pickle file to perform prediction on Patient data.

The trained Random Forest algorithm for kidney disease prediction and Support Vector Machine algorithm for Liver disease prediction is used. All other lab test values which are higher or lower than the reference value are highlighted with red color to show the abnormal flag so that the doctor will focus on abnormal values and can easily scan the lab test result in less time. The whole process is executed in Jupyter Notebook[19] and Python Sklearn[20] library is used for machine learning algorithm implementation, and precision, accuracy, recall and f-score calculation.

## C. Patient Identification Using Face Recognition

The system uses a biometric-based patient identification for the misidentification issue. The face recognition system helps in identifying the patient in real-time and access the patient medical record instantly. The implemented system aimed at reducing the patient's medical record access time.

The proposed system requires minimal changes in the traditional system. A webcam attached to a computer or mobile phone can be used to identify the patient instantly. It requires a picture of the patient for registration and identification. When patient identification is performed for the first time, the system checks for the patient's face in a registered patient database. If a face is not registered, a Unique Health Identity (UHID) is assigned to the patient. The system doesn't store a patient face image, instead of an image, face feature encoding of 128 points is stored in the database. The ‘face_recognition’[15] library is used to generate encoding and face detection. The face_recognition[15] library uses deep learning library dlib[14] for face feature encoding and histogram of oriented gradients(HOG) [21] for face detection.

## D. Python Django Server

An open-source web framework Django[22] is used to develop the REST APIs for Face Recognition and Clinical Decision Support System (CDSS). The Django uses PostgreSQL as a database. All the machine learning model and dataset are served as web services using Django HTTP interface.

Get List, Create, Update, Delete, and Read REST APIs for User, OPD, Hospital, Doctor, Patient, Face Recognition, Prediction, Lab, Prescription modules are developed and secured by token-based authentication using email and password.

## E. React.JS Web Application

The clinical decision support system web-based User Interface was developed in React.JS. React is a javascript library developed by facebook for developing a single-page application with component style. The web app consists of OPD, Prescription, Patient, Doctor, Hospital Staff, and Lab module. The clinical decision support system module was developed in web application and integrate it into the prescription creation form to ease the usage for the doctor. REST APIs are used to interact with the backend Django Server.

Fig. 5 shows the Disease Prediction module in React Application. It uses the autocomplete dropdown for disease selection. On click the predict button, the Application makes the request to the backend for prediction. Backend returns the list of diseases to UI which is shown below the predict button.

Fig. 6 shows a list of suggestions given by the prediction module. It shows the recommended lab test, medication for GERD to that the doctor can prescribe the correct treatment.

Fig. 7 shows the chronic kidney prediction module. It auto-fills the details in the field by analyzing the lab reports of the patient. At the bottom of the module, it shows the probability of chronic kidney disease in terms of percentage.

Fig. 8 shows the liver disease prediction module which is used to check the risk of liver disease. It shows the probability of liver disease in terms of percentage.

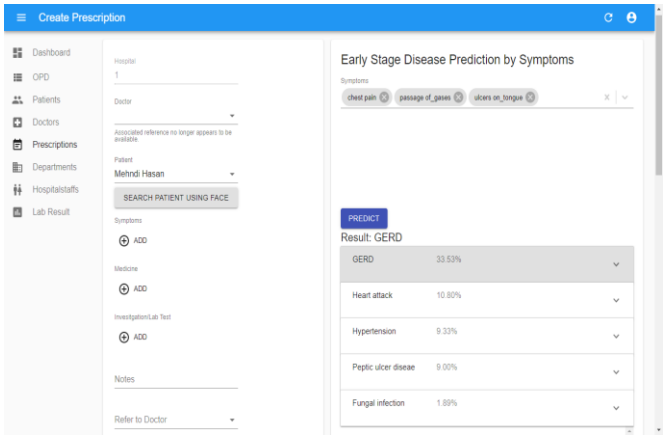


**Fig 5. Disease Prediction Module**



Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

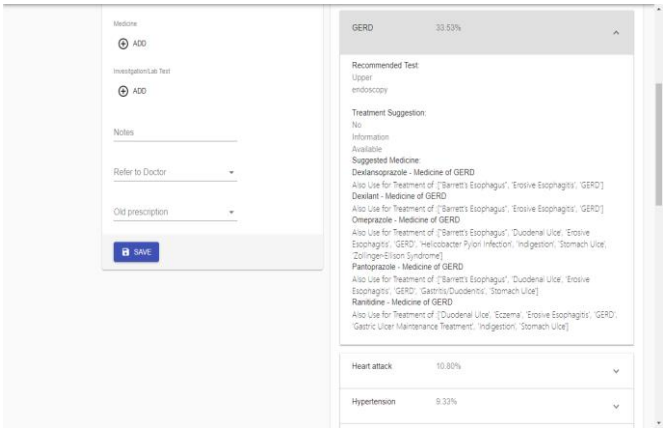


**Fig. 6. Treatment Suggestion in Disease Prediction Module**

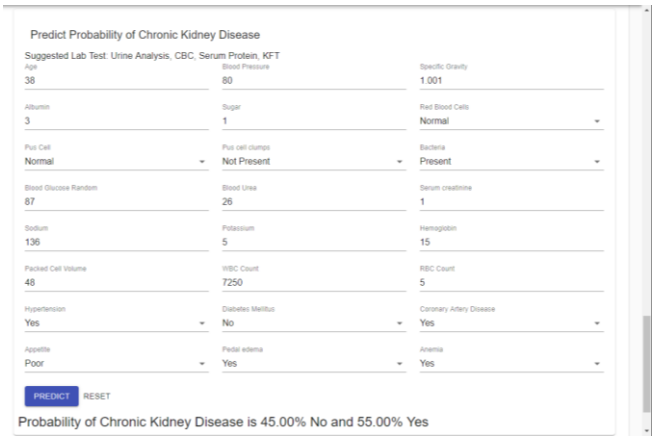


**Fig. 7. Chronic Kidney Disease Prediction Module**


**Figure 8: Liver Disease Prediction Module**

## IV. EXPERIMENTAL RESULT

### A. Hardware and Software Used

All the Softwares used are listed below:

- PostgreSQL: An open-source relational database engine
- PyCharm: a Python IDE software.
- WebStorm: an IDE for Web Development
- Zotero: a citation management tool
- Chrome Browser
- Google Docs: for documentation
- Google Drawing: for making flowcharts
- Grammarly: for checking grammatical mistakes
- Jupyter: an open-source interactive console for Python and other languages.

The Hardware used are listed below:

- Laptop i3 5th gen, 8 GB RAM, 1TB HDD, Windows 10

### B. Parameter of Evaluation

Before defining the evaluation parameter, the meaning of a few terms should be clear.

- **TP** - True positive samples (TP) are samples that were classified positive and are really positive.
- **FP** - False positive samples (FP) are samples that were classified positive but should have been classified negative.
- **FN** - False-negative samples (FN) were classified as negative but should be positive.

#### 1) Precision

$$P = \frac{TP}{TP+FP}$$

Precision (P) can be intuitively understood as the classifier's ability to only predict really positive samples as positive [23]. For example, a classifier that classifies just everything as positive would have a precision of 0.5 in a balanced test set (50% positive, 50% negative). One that has no false positives i.e. classifies only the true positives as positive would have a precision of 1.0. So basically, the less false positives a classifier gives, the higher is its precision.

#### 2) Recall

$$R = \frac{TP}{TP+FN}$$

Recall (R) can be interpreted as the number of positive test samples that were actually classified as positive [23]. A classifier that just outputs positive for every sample, regardless if it is really positive, would get a recall of 1.0 but lower precision. The less false negatives a classifier gives, the higher is its recall.

#### 3) Accuracy

$$A = \frac{Total\ Positive}{Total\ Samples}$$

As a heuristic, or rule of thumb, accuracy can tell us immediately whether a model is being trained correctly and how it may perform generally. However, it does not give detailed information regarding its application to the problem.[24]

#### 4) F-score

$$F1 = 2\,\frac{P*R}{P+R}$$

This is just the weighted average between precision and recall. F-score is the harmonic mean of precision and recall. The higher precision and recall are, the higher the F1-score or F-score is.[24]

### C. Comparison of Various Machine Learning Algorithms used on Chronic Kidney Disease Dataset

Performance analysis of machine learning algorithms for kidney disease diagnosis is performed.

Preprocessing and implementation steps are discussed in section 3.2.2 (a) of the paper. We have selected the most used algorithms i.e. Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and K-nearest Neighbor (KNN) algorithm for comparison.

**Table-II: Comparison of Various Classification Algorithms for Kidney Disease Prediction.**

| Algo | Accuracy | P | R | F1 |
|---|---|---|---|---|
| **SVM** | 0.72 ( 72.91% ) | 0.73 | 1.0 | 0.84 |
| **LR** | 0.97 (97.91 % ) | 0.97 | 1.0 | 0.98 |
| **KNN** | 0.81 (81.25%) | 0.79 | 1.0 | 0.88 |
| **RF** | 1.0 ( 100% ) | 1.0 | 1.0 | 1.0 |

Table-II shows the performance comparison of SVM, LR, KNN and RF in terms of accuracy score, precision, recall, and f-score. Fig. 9 shows the graphical representation of the accuracy score, precision, recall and f-score of SVM, LR, and RF. The x-axis represents the various classification algorithms and the y-axis represents the accuracy score, f-score, precision and recall of algorithms.

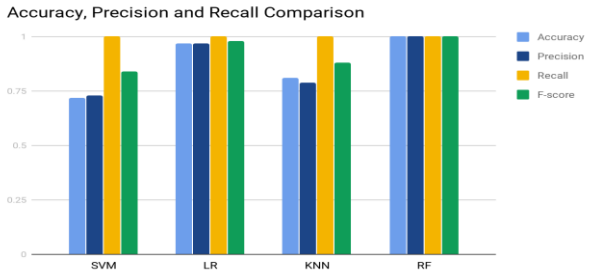


**Fig. 9. Graphical Comparison of Various Algorithms for Kidney Disease Prediction**

From the above results, it is clear that Random Forest outperforms the other algorithms. We have selected the Random Forest algorithm for kidney disease prediction.

### D. Comparison Of Various Classifier Algorithms Used On Liver Disease Dataset

Performance analysis of machine learning algorithms for liver disease prediction is performed. The details about the dataset and implementation steps are discussed in section 3.2.2 (b) of the paper. We have used Support Vector Machine (SVM), Random Forest (RF), K-nearest Neighbor (KNN) and Logistic Regression (LR) algorithms for comparison and performance analysis.

**Table-III. Comparison of Various Classification Algorithms for Liver Disease Prediction.**

| Algo | Accuracy Score | P | R | F1 |
|---|---|---|---|---|
| **LR** | 0.64 (64.36 %) | 0.69 | 0.88 | 0.77 |
| **SVM** | 0.70 (70.68 %) | 0.70 | 1.0 | 0.82 |
| **KNN** | 0.61 (61.50 %) | 0.71 | 0.75 | 0.73 |
| **RF** | 0.68 (68.96 %) | 0.73 | 0.86 | 0.77 |

Table-III shows the performance comparison of SVM, LR, KNN, and RF in terms of accuracy score, precision, recall, and f-score. Fig. 10 shows the graphical representation of the accuracy score, precision, recall and f-score of SVM, LR, KNN, and RF. The x-axis represents the various classification algorithms and the y-axis represents the accuracy score, f-score, precision and recall of algorithms.

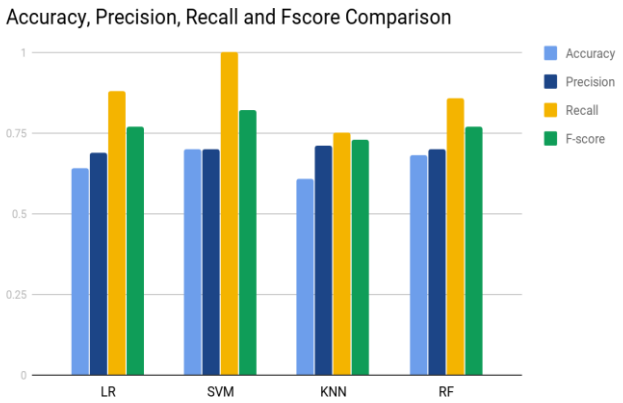


**Fig. 10. Accuracy Score of Various Classifiers for Liver Disease Prediction.**

Support Vector Machine has the highest accuracy (70%) among all. We have selected the Support Vector Machine algorithm for liver disease prediction.

### E. Comparison of the Clinical Decision Support System Prediction with Doctors Analysis

As Clinical Decision Support System (CDSS) is developed to assist doctors not to compete with the doctor. CDSS provides the information to the doctor about all the possible diseases and treatment plans of those diseases.

First Author has collected symptoms and prescription data of patients in OPD at a Siddhanta Redcross Superspeciality Hospital[25] in Bhopal (M.P). We have compared the Doctor's outcome and CDSS suggestions on the basis of symptoms. If the proposed system is able to suggest the disease than the author has considered it positive outcome else negative

**Case #1: Patient 1 is 75 Years Old Female**

Initially, she had a fever, vomiting, stomach pain, fatigue, and blood pressure issues. The doctor suggested Laparoscopy (Endoscopy), LFT, KFT, HbA1C to diagnose the disease. After the report of Laparoscopy. She has diagnosed cholelithiasis (gallstones in the common bile duct) and admitted to the hospital. We provide the same symptoms in CDSS as input, CDSS Recommendation was chronic cholestasis (Disorders of the liver, bile duct, or pancreas) with a 22% probability at first position out of 5 suggestions. Both diseases are related to bile duct and stone.

**Outcome: POSITIVE**

**Case #2: Patient 2 is 43 Years Old Male**

He was having chest pain and gastric issue for a long time. The doctor suggested UGIE + RUT (Endoscopy) and medication for the gastric issue. After the report of UGIE & RUT, Rapid Urease Test was Positive which means Ulcer in the stomach. The same symptoms are provided to CDSS, CDSS suggested Peptic ulcer disease with 11.30% probability at third out of 5 suggestions.

**Outcome: POSITIVE**

**Case #3: Patient 3 is 43 Years Old Male**

He was suffering from gastric problems, stiffness in abdominal, and stomach pain. The doctor suggested UGIE + RUT, LFT, CBP, USG Abdominal lab test and initial observation was GERD.

After lab reports, He had GERD, Fatty Liver, dyslipidemia. Same symptoms provided to CDSS, System recommended GERD at 3rd Position out of 5 with 11.89% probability. We have collected lab test reports of patient and performed lab test analysis on it.

Liver Disease Prediction: Probability of Liver Disease is **60.00% Yes** and 40.00% No

Kidney Disease Prediction: Probability of Chronic Kidney Disease is 39.00% Yes & 61.00% No

**Outcome: POSITIVE**

**Case #4: Patient 4 is 46 Years Old Female**

She was suffering from upper abdominal pain, gas issues, vomiting, nausea, appetite low. The doctor suggested UGIE + RUT, CBP, Urine Culture, Random Sugar, LFT, T3, T4, TSH tests and prescription medication for the gastric issue and abdominal pain. After the report of UGIE, Rapid Urease Test was Positive, She was diagnosed with a small ulcer in the stomach. The suggestion given by CDSS was Peptic ulcer disease with 39.67% probability at first position out of 5 suggestions.

**Outcome: Positive**

**Case #5: Patient 5 is 38 Years Old Male**

He was suffering from vomiting, loss of appetite, stomach burn. He had a previous medical history of TB, Ulcer, and Crohn's disease. The doctor suggested the Mantoux test, ESR, CRP, CBC, KFT, LFT, and Random Sugar with endoscopy of the gastric outlet. His endoscopy shows Gastro-jejunostomy and ulcers in the stomach. We provide the same symptoms to CDSS. The suggestion given by CDSS was GERD (with 22.71% probability at first position) followed by Peptic Ulcer Disease (with 14.63% probability at the second position)

Probability of Chronic Kidney Disease is 57.00% No and 43.00% Yes

Probability of Liver Disease is 49.00% No and 51.00% Yes

**Outcome: POSITIVE**

Table – IV shows the comparison of doctor's outcome and proposed data-driven clinical decision support system. The data-driven clinical decision support system correctly predicts the possibility of the disease in the above-discussed patient cases. In 60% (3 out of 5 cases) cases, the proposed system suggests the disease at top places with the probability of more than 20%. In all the cases, the proposed system suggested the disease name in the top 5 suggestions. The result of the proposed system shows the accuracy of the machine learning-based clinical decision support system in the diagnosis of possible disease.

## V. CONCLUSION AND FUTURE WORK

The proposed and implemented system aimed to diagnose the disease at an early stage and provide suggestions to the doctor to improve healthcare. This kind of system can show the nearest possibility of disease that a doctor might not have thought of while diagnosing. The system is designed to work on minimal input by the doctor so that the doctor can use the system in OPD by simply providing the symptoms of the patient. It also provides the features of integration with EHR and provides REST API that can be used in a mobile app to build a handy CDSS. Thus, this system can be available to all the doctors of India through their smartphone.

Developed CDSS needs a minimal change in the workflow of the doctor but in the long run, it will create a huge impact in improving the care provided to patients by diagnosing all the possibility at an early stage. The implemented system will improve the possibility of disease by analyzing the lab results. Currently, Lab test analyzer predicts liver disease and kidney disease. In the future, more such types of modules can be developed to increase the scope of lab test analysis.

## REFERENCES


1. "What is an electronic health record (EHR)? | HealthIT.gov." [Online]. Available: https://www.healthit.gov/faq/what-electronic-health-record-ehr. [Accessed: 17-Jul-2019].
2. "What is clinical decision support system (CDSS)? - Definition from WhatIs.com," SearchHealthIT. [Online]. Available: https://searchhealthit.techtarget.com/definition/clinical-decision-support-system-CDSS. [Accessed: 19-Aug-2019].
3. "UCI Machine Learning Repository: Drug Review Dataset (Drugs.com) Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29. [Accessed: 19-Aug-2019].
4. "What are Clinical Rules? | Clinical Rules." [Online]. Available: http://www.clinicalrules.nl/en/what-is-clinical-rules. [Accessed: 19-Aug-2019].
5. "Guide your patients to the right care – Infermedica." [Online]. Available: https://infermedica.com/. [Accessed: 19-Aug-2019].
6. "Symptom Checker - Esagil. Check Medical Symptoms." [Online]. Available: http://esagil.com/. [Accessed: 19-Aug-2019].
7. "Isabel Healthcare | Differential Diagnosis Tool." [Online]. Available: https://www.isabelhealthcare.com/. [Accessed: 19-Aug-2019].
8. "About Products," SimulConsult. [Online]. Available: https://simulconsult.com/about-products/. [Accessed: 19-Aug-2019].
9. "DISEASE PREDICTION DATASET" [Online]. Available: https://kaggle.com/neelima98/disease-prediction-using-machine-learning. [Accessed: 19-Aug-2019].
10. "Datasets | Kaggle." [Online]. Available: https://www.kaggle.com/datasets. [Accessed: 19-Aug-2019].
11. "UCI Machine Learning Repository: ILPD (Indian Liver Patient Dataset) Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29. [Accessed: 31-Aug-2019].


**Table- IV: Comparison of CDSS Disease Prediction with Doctor Diagnosis**

| S.No | Doctor Outcome | CDSS Outcome | Probability of Prediction | Position of Suggestion | Outcome |
|---|---|---|---|---|---|
| 1 | Gallstone in the bile duct | Disorders of the liver, bile duct | 22% | 1st | Positive |
| 2 | Ulcer | Peptic ulcer disease | 11.30% | 3rd | Positive |
| 3 | GERD, Fatty Liver, dyslipidemia | GERD | 11.89% | 3rd | Positive |
| 4 | Small Ulcer | Peptic ulcer disease | 39.67% | 1st | Positive |
| 5 | Gastro-jejunostomy and ulcer | GERD & Peptic Ulcer Disease | 22.71% & 14.63% | 1st and 2nd | Positive |



Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

12. "UCI Machine Learning Repository." [Online]. Available: https://archive.ics.uci.edu/ml/index.php. [Accessed: 19-Aug-2019].
13. "UCI Machine Learning Repository: Chronic_Kidney_Disease Data Set." [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease#. [Accessed: 19-Aug-2019].
14. "dlib C++ Library." [Online]. Available: http://dlib.net/. [Accessed: 26-Aug-2019].
15. A. Geitgey, The world's simplest facial recognition api for Python and the command line: ageitgey/face_recognition. 2019 [Online]. Availale: https://github.com/ageitgey/face_recognition/.
16. "WebMD - Better information. Better health.," WebMD. [Online]. Available: https://www.webmd.com/default.htm. [Accessed: 19-Aug-2019].
17. "Access Anytime Anywhere," Cleveland Clinic. [Online]. Available: https://my.clevelandclinic.org/. [Accessed: 19-Aug-2019].
18. "UCSF Medical Center." [Online]. Available: https://www.ucsfhealth.org/. [Accessed: 19-Aug-2019].
19. "Project Jupyter." [Online]. Available: https://www.jupyter.org. [Accessed: 06-Sep-2019].
20. "scikit-learn: machine learning in Python — scikit-learn 0.21.3 documentation." [Online]. Available: https://scikit-learn.org/stable/. [Accessed: 06-Sep-2019].
21. "Histogram of oriented gradients," Wikipedia. 03-Apr-2019.
22. "The Web framework for perfectionists with deadlines | Django." [Online]. Available: https://www.djangoproject.com/. [Accessed: 09-Jul-2019].
23. N. Arora and M. Motwani, "Optimizing K-Means by Fixing Initial Cluster Centers," Int. J. Curr. Eng. Technol., vol. 4, no. 3, pp. 2101–2107, Jan. 2011.
24. "Evaluation Metrics for Machine Learning - Accuracy, Precision, Recall, and F1 Defined," Skymind. [Online]. Available: http://skymind.ai/wiki/accuracy-precision-recall-f1. [Accessed: 31-Aug-2019].
25. "Super Speciality Hospital in Bhopal | Plastic Surgery in Bhopal | Cardiology in Bhopal." [Online]. Available: http://www.siddhantaredcrosshospital.com/. [Accessed: 03-Sep-2019].

## AUTHORS PROFILE

**Mr. Shubham Rathi** is currently pursuing Dual Degree Integrated Post Graduate Programme (B.E + M.Tech) in Computer Science and Engineering from University Institute of Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal (MP), India. His research areas are machine learning, health technology, and blockchain. He has won National and State level competition for his Innovative Project 'WiBin'.

**Dr. Mahesh Motwani** is Professor in Department of Computer Science & Engineering at University Institute of Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal (MP), India. He has around 28 Years of teaching experience and around 1 Year of experience as a Scientist at National Informatics Centre (NIC), Delhi. His areas of research are data mining, ad-hoc network, and machine learning. He has published many research papers in National and International Journals and Conferences. 6 Ph.D. have been awarded under his guidance.

**Mr. Manish Kumar Ahirwar** is currently working as an Assistant Professor in Department of Computer Science, University Institute of Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal (MP), India. He has work experience of several years in the field of teaching. His research interests include Data Mining Algorithms, Internet of Things (IoT), Machine Learning and Cyber Security. He is a member of IEEE, ACM, IACSIT, IAENG. He has published more than 30 research papers in various International and National Journals and Conferences, including 4 papers in SCIE Journals and more than 10 papers in Scopus Journals. He has also published two Indian patents and two copyrights.