Cognitive Neurodynamics (2022) 16:1203–1220
https://doi.org/10.1007/s11571-021-09758-y

RESEARCH ARTICLE

An ensemble approach for healthcare application and diagnosis using natural language processing

Badi Alekhya[1] · R. Sasikumar[1]

Received: 20 September 2021 / Revised: 7 November 2021 / Accepted: 22 November 2021 / Published online: 17 January 2022

Abstract

Integration of healthcare records into a single application is still a challenging process There are additional issues when data becomes heterogeneous, and its application based on users does not appear to be the same. Hence, we propose an application called MEDSHARE which is a web-based application that integrates the data from various sources and helps the patient to access all their health records in a single point of source. Apart just from the collection of data, this portal enables the process of diagnosis using Natural language processing. The process is carried out by fuzzy logic ruleset which is generated by using NLP packages. The resulted information is given to the SVM classifier which helps in the prediction of diseases resulting in 89% of accuracy and standing the best compared to other classifiers. Finally, the observations resulted are sent to the front end application and the concerned user mobile through text message in their own native language for which translation package is been used.

Keywords Heterogeneous data integration · MEDSHARE · Natural language processing · SVM · Fuzzy logic

Introduction

There are a variety of traditional healthcare support systems available that integrate heterogeneous information (Popowich 2005; Névéol and Zweigenbaum 2015). These systems, however, are prone to a myriad of issues such as data quality, sparseness, ambiguous information, etc. Processing invalid or ambiguous information leads to faulty analytics of faulty predictions. In the healthcare field, faulty predictions may lead to a loss of life. Hence, a novel system that can identify chronic diseases accurately needs to be found. When considering chronic diseases it is one of the main threats to mankind and seems to be challenging to the healthcare systems found around the world. Also, when we consider population, it is rapidly expanding on a daily basis, which adds another issue. Given today's population, the population appears to exceed 21% by 2050.

Hence to support this large population our medical systems need more advancement (Kaur 2020) to cater to the needs of people from a rural area to topmost management. As a result, the existing system in healthcare needs to be enhanced that enables to rectification of the gap present such as the shortage of resources, efficiency, and cost. Automatic or remote access to medical information or diagnoses is becoming increasingly popular since it is more practicable, cost-effective, and reliable. Healthcare organizations are developing a myriad of applications to handle all of these capabilities while also catering to the needs of their users' environments. Due to more advancement in technologies, there is generally a focus towards the process of expert medical systems to be adopted that supports diagnosis and treatment (Srinivasan and Madheswari 2017, 2018; Gowthul Alam and Baulkani 2017, 2019a, b; Nanjappan and Albert 2019; Nanjappan et al. 2021).

The diagnostic process includes technology related to the computer which is improved over the period of time thus making physicians help in accurate diagnosis of disease by enabling the signal processing techniques, association rule mining algorithms, and neural networks on the process of making decisions. Thus the patient must be able to get involved in keep track of his own health and must be more involved as a key person in managing and taking decisions together with the providers of healthcare (Coulter and Collins 2011; Dick 1997). As a result, the methods and

✉ Badi Alekhya
alekhyareddy1@gmail.com

[1] R. M. D. Engineering College, Kavaraipettai, Chennai, India

developments made to these techniques must be more viable in the case of self-management and should have a proper interface for communication (Haseena et al. 2014; Kavitha and Ravikumar 2021; Rejeesh and Thejaswini 2020, Ravikumar and Kavitha 2020, 2021; Sundararaj et al. 2020; Sundararaj 2019; Sundararaj and Selvi 2021; Vinu and Sundararaj 2016; Xie et al. 2021).

There are a large number of healthcare providers that help in maintaining the Electronic Health Records (EHR) (Légaré and Holly 2013; Walshe and Rundall 2001; Ash et al. 2004). The development approach should be adaptable to portable devices, and a self-monitoring report in his/her local language with fundamental parameters (Li et al. 2019; Wu et al. 1990) in an understandable style should be enabled. Thus, implementing this type of application will involve patients in taking care of their health, resulting in the collection of a substantial quantity of health-related information. When bringing in this application, it seems that there are so many techniques that are used to collect health information. Wearable devices are used to monitor the heartbeat rate, pulse, and other health data regularly to maintain a healthy lifestyle. Similarly, additional future and advanced applications for health maintenance have raised the concept of self-monitoring as a practical option for patients. Thus, capturing all of this information aids in the accumulation of massive amounts of data, which are then used for a variety of prediction and analysis purposes (Bodenreider 2004; Ghasemzadeh et al. 2012; Ravikumar and Kavitha 2021; Walczak 2005).

Hence the collaborative amount of high-end data about the patient will enable and provide potential and valuable input to the providers of healthcare while decision making. These all are the essence of the development of data mining in healthcare, clinical decision support making systems, and so on (Hardin and Chhieng 2007; Zhang et al. 2012). Conversely, because the application's contribution to these data is rather heterogeneous, the system's connectivity and accessing nature is discovered to be diverse when it is being built. The main disadvantage of focusing on this heterogeneous nature is that efficiency will be difficult to achieve because data utilization and transmission are the most important factors. To overcome this, a data model (Corral-Plaza et al. 2020) must be built-in that supports and provides a solution to all of these types of problems.

This scenario needs more integration methods that enable the combination of distributed data related to healthcare from numerous systems to be made available in a single platform. This can support different devices including mobile, framework-based windows, semantic web, and other unexplored devices. To be specific there is no existing literature study that supports all these features and still, it's not highlighted under a single specific topic. Natural Language Processing(NLP) (Stenetorp et al. 2012) help in acting as an interface between humans and computer so that they can interact. It makes use of linguistic analysis which means that in spite of language there are options available that can help the system to work in any sort of language category. The paper presents a novel MEDical related Secure Heterogeneous data-based Application and REaltime disease diagnosis (MEDSHARE) system with the following features.

- The primary goal of this study is to integrate patients' heterogeneous medical records and deliver disease prediction results in their native language.
- The MEDSHARE application can be operated by both the patient and the medical institution staff.
- A two-way authentication scheme is provided to ensure the security of the users.
- The SVM classifier is incorporated to predict the disease of the patients using the blood test results with the help of the fuzzy logic module.
- The results obtained show that the fuzzy logic module integrated with the SVM classifier yield improved accuracy.

The remainder of the paper is structured accordingly. The entire paper involves the following sections: "Review of related works" section provides a survey of several healthcare applications for integrating data and heterogeneous data. Followed by that in "Proposed approach" section the proposed methodology is discussed and the results obtained are given in "Results" section. Finally, the conclusion and future work are given in 'Conclusion" section.

Review of related works

This section includes a detailed investigation of the existing techniques that are present which are involved in the integration of healthcare data. They don't exactly provide the same solution that we have proposed in this paper yet are still able to check the challenges that are present in this field and help to bring up collaborative healthcare data utilization. Main challenges are highlighted which include the utilization of data and integration further added feature will be the efficiency of data.

Srivastava and Singh (2020) presented a Medibot technique to identify the user's symptoms with a 65% precision via NLP. Kandpal et al. (2020) presented a Contextual chatbot to offer healthcare services using deep learning. However, these applications can be only operated in the English language. Jungmann et al. (2020) presented an NLP technique to identify the patient suspected of urolithiasis by extracting the information automatically

from the EHR. However, they haven't analyzed the performance of this model. Kaur et al. (2018) automatically extracted the Asthma Prediction Index(API) from the user's EHR via NLP to identify the patients that mark the API criterion. The API mainly denotes the risk factor for asthma patients. Zhong et al. (2019) utilized the NLP technique to identify pregnant women with suicidal behavior. However, when compared to the data obtained using diagnosis codes, the use of unstructured text revealed that there are 11 times as many pregnant women with suicidal behavior.

The deep learning methods (Shickel et al. 2017) are primarily focused and used so that knowledge can be extracted from the free unstructured text. These further prove they are exceptional and are mandatory in the region of information retrieval and also a large amount of unstructured data to be processed. Liu et al. (2019) proposed an ensemble of NLP systems for automatic phenotype extraction from clinical texts. They have evaluated the performance of their technique by analyzing generic phenotypic concepts and patient-specific phenotypic concepts. They created different types of ensemble architectures such as majority vote-based ensemble, training-based ensemble, union ensemble, MetaMapLite, MedLEE, and cTAKES. They mainly proved that integrating multiple ensemble NLP techniques can improve the performance and reproducibility of the results. However, the training model should be altered for each performance metric to reach certain objectives.

Tvardik et al. (2018) evaluated the performance of the SYNODOS NLP solution in identifying healthcare-associated infections. Different medical fields such as adult intensive care unit, digestive surgery, neurosurgery, and orthopedic surgery were involved in this study. The accuracy rate, sensitivity, and specificity value for their technique in identifying healthcare-associated infections is 84%, 83.9%, and 84.2% respectively. Solomon et al. (2021) used NLP for aortic stenosis identification and its severity calculation. The authors found that the validated NLP algorithm yielded more accuracy for the electrocardiography database than the diagnosis codes. The administrative diagnosis codes and the operating nature of the NLP algorithms in diagnosing aortic stenosis, however, have considerable discrepancies.

Habib et al. (2021) presented a methodology known as AltibbiVec which is a neural-based word embedding model designed in the Arabic language for medical and healthcare applications. The AltibbiVec architecture identifies the appropriate and useful features from the thousands of data samples. They have used three different embedding models namely GloVe, fastText, and Word2Vec. Based on their results, the Word2Vec and fastText were more effective when compared to GloVe. However, their application requires abundant data for contextualized word representations. But in spite of having these many advantages the main challenge is the formatting process of healthcare records, keywords not available, use of so many punctuations, not correct parts of speech, spelling mistakes, etc. The use of linguistic analysis (Roberts 2017) finds tough as the text containing medical terms is difficult to interpret. Sometimes instead of full-text abbreviations being used it becomes even harder to come up with inference (Yu et al. 2020). Hence due to this sometimes researchers find challenging in using NLP, especially for medical domains. To overcome these problems in our research we are proposing an application that helps in categorizing before coming to inference using fuzzy logic rules and word embedding technique. This technique also sends the prediction results to the users via an SMS in their native language because in India people do not always have an uninterrupted network connection.

Proposed approach

Data integration of healthcare records-related techniques and terminologies are still a new topic and are emerging still in multi-dimensions. Before going on to the proposed technique, here the main term data integration includes a combination of health care data through distributed vendors and sources that are organized under a single application and seem to be coming from a single source to patients. Source of data can include clinical labs, scan centers, hospitals, and devices to maintain health such as healthcare record systems and so on. Hence our proposed application will help in collaborating all these data as a single unit. Figure 1 explains the workflow of the proposed methodology. The data integration technique along with its utilization can seem to be bound with each other. The integration of data in this paper includes a collection of data from various sources added to that other features are also included which are listed as follows, (1) Data collection and aggregation, (2) Medical conclusion, (3) Preference in the native language, and (iv) Diagnosis and prediction process.

The overall process includes integration, utilization, processing, classification, and visualization of data is been described in Fig. 1. The data can be collected from various sources as stated earlier. Hence it is represented as service providers and each service provider will not collect data from one source instead they will be collecting it through multiple sources. Further, all providers giving data can be integrated and collected which proceeded with a heterogeneous process. After the heterogeneous data integration process is completed, the knowledge is extracted from the data using different techniques such as UMLS, WordNet,

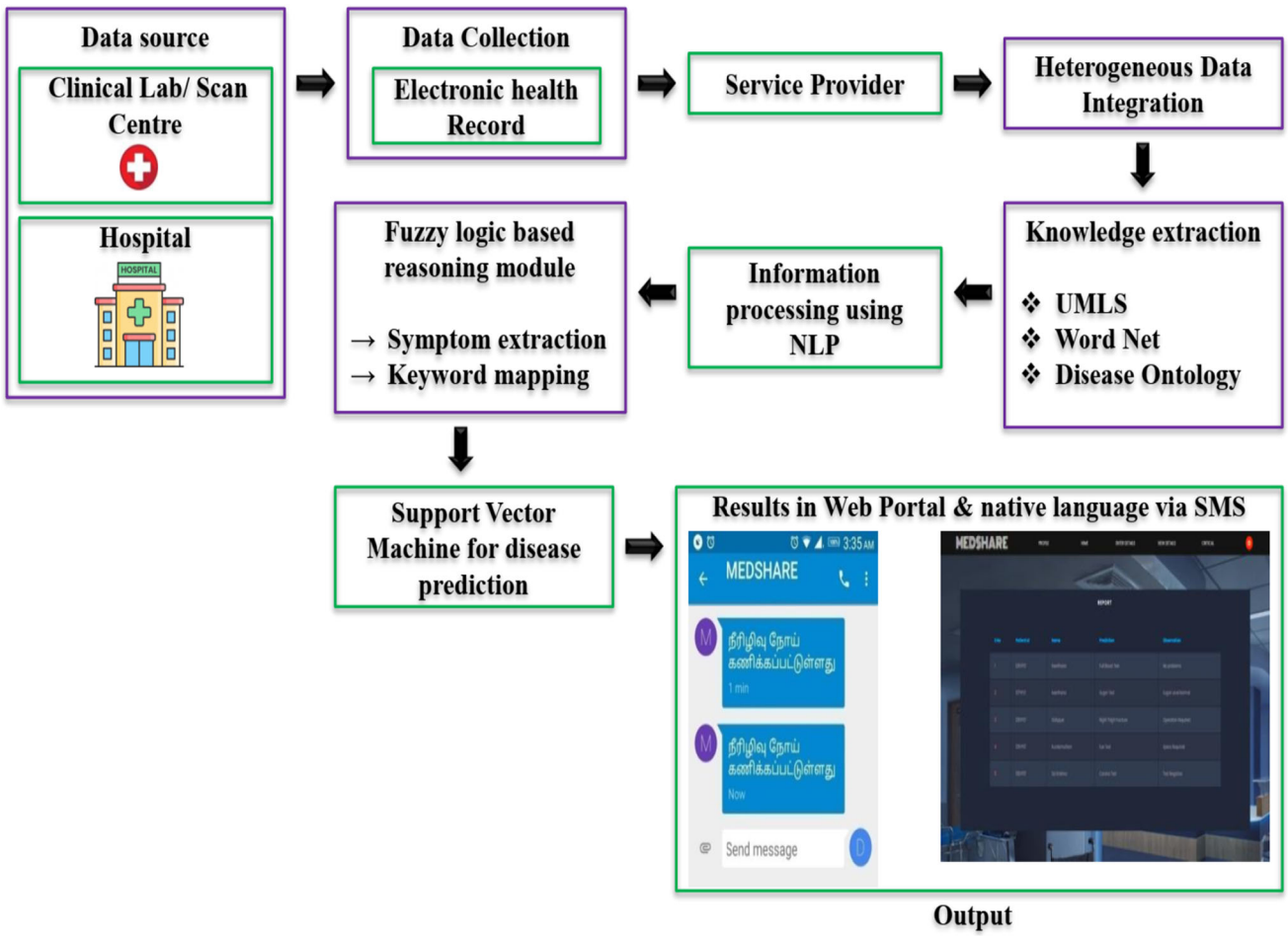

Fig. 1 Workflow of the proposed technique

and disease ontology. The NLP extracts the crucial features for disease diagnosis and the results are given the SVM classifier and the fuzzy logic-based reasoning module. Using symptom extraction and keyword mapping, the fuzzy rule generation framework identifies the patients who are prone to multiple diseases. The SVM classifier classifies the disease of the patients based on the features extracted. The data visualization process helps the patients to witness the results and it can be finally used for the prediction of any disease in the near future.

Alongside the term of data integration, another well-defined feature is needed which is known to be interoperability. This is said to be related very closely to the integration of data because of a different perspective. The interoperability in terms of healthcare can be considered at three levels: At the base-foundation level, Structural and Semantic. As the name goes the foundation level of operability includes the data to be exchanged from one system to another and the definition of data along different classes can be said to be structural level. It is termed to be the ability of data to be interpreted through the syntax of data. The third classification is said to be the peak level which includes coding the data from the previous level that is from syntax converted so that system will be able to interpret and come up with conclusions.

During this process when a certain level is achieved then these healthcare data can be transferred to any devices, viewed in terms of graphs or data, can be able to see interpreted observations, and can be utilized by anyone who needs it. Hence this operation is said to be a prerequisite for the integration of data. Thus these additional features are needed to develop the proposed applications. The steps involved in the proposed architecture are presented in Fig. 2. The overall architecture in includes (1) front-end web application that collects the data from different service providers; (2) extraction of Keyword and relevant information; (3) prediction and diagnosis system; (4) result at web and mobile platform; and (5) result at preferred language. The entire implementation process is carried out using python since it contains many advanced packages that support machine learning and Natural language processing.

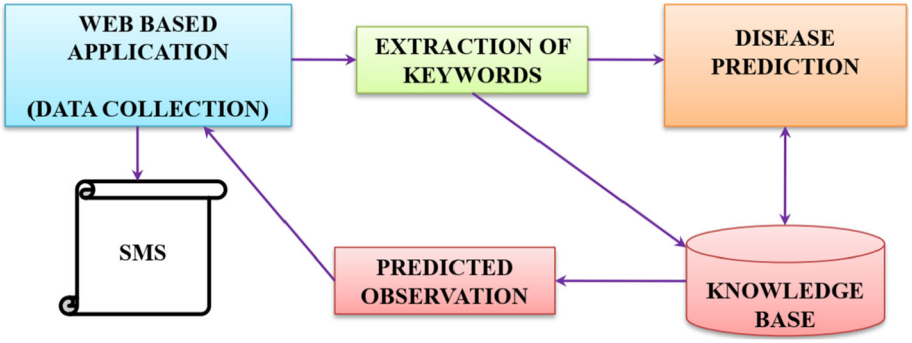

Fig. 2 Steps involved in the proposed architecture

Web-based application

The front end for user communication and other service providers' communication is collected through web sources. This application is categorized into two different parts. One for patients' views and the other for service providers such as clinical labs, scan centers, hospitals, and other service providers. As the same goes for all applications this includes the process of registering based on radar identity and other basic information. The functionalities include the following features: login, register, view reports, and get observations status. The next category at the other end includes: login, register, view the critical status of patients, add patient info, and edit the health record. So through this, the patients at any end will be able to know their health status as prescribed by the doctors and will be able to communicate it in case of other higher-end treatments.

The data can be manipulated or accessed by hackers which in that case makes the data to be unreliable. Hence proper authentication is provided so as to secure the data. Here the login process can be done only through mobile OTP which then makes only the user login. Additional security features can be provided which is again the vast area. Hence here we are focusing on the front end part to collect and view the information at various ends. The application seems to be secure in case of handling with a one-time password as the other users are not able to access it without the code. Two-way authentication (Kothmayr et al. 2013) is provided at the other end say as to be hospital side where double verification system procedure is carried out. This is done here as they must be very reliable since the overall diagnosis and prediction system will be working based on the data entered by them added to which the patient/user might go with it and believe.

Information extraction

This is the second part and foremost part of this application. It becomes the principal component of the proposed system. It includes various other data to be stored in form of structural and unstructured forms. The data present here is categorized and collected so that the prediction process and diagnosis process will be able to do under different disease contexts. It contains the following things to be integrated that support as the base of the natural language processing.

1. *UMLS (Unified Medical Language System)*—It combines vocabularies related to medical terms and is further used to map the terms with original medical facts and knowledge. This is considered to be one of the best systems that are considered to be a standard knowledge source in the domain of healthcare (Bodenreider, Olivier, 2004). Implemented using UMLS REST API in python language. With help of import entire methods are used from this package.
2. *WordNet* (Morato and Marzal 2004)—Since this application includes a collection of data in various forms this database is used. It is a lexical database that helps in defining the relationship between phrases and words. This is a familiar repository in dealing with words. This can be implemented by using API.
3. *Disease Ontology (D6)* (Schriml et al. 2012)—This is used as it contains information on more than 10,000 human diseases. As the same as UMLS here metadata can be accessed by adding one more term DO to the same REST API further sending a request.

Hence this entire process of knowledge bed is carried out through XML (Extensible Markup Language) through which the data are processed in and out of the knowledge bed. The entire data given at the observation part are categorized into three levels. All the healthcare data includes this process: a collection of properties, symptoms, and disease prediction. The categorization of data is depicted in Fig. 3.

Extraction of information

The overall data can be considered at three-level as stated in the above part. The basic information includes the name, id of the patient, dob, and other related information about the patient. These can be considered as the properties of

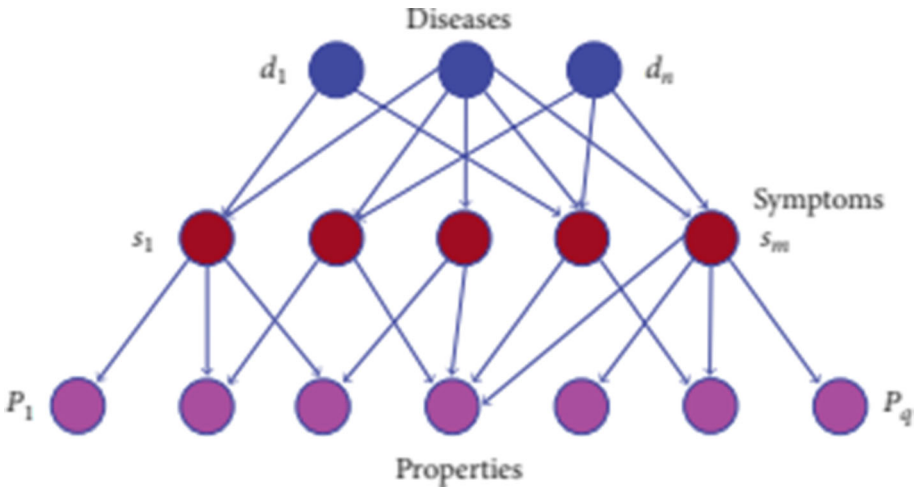

Fig. 3 Categorization of data

that particular patient. The patient id is considered to be the primary key through which the other data are retrieved. The weight, age, and birthmark all are considered to be the primary information. The extraction of this information and categorization is done and explained in Algorithm 1.

Hence when the value crosses the threshold value as mentioned in sets used then the disease name is mapped with the id created. Added to this the doctor observation is been checked as they are the final authorized persons to be confirmed and approval of diseases. The algorithm can be

```
def data_extraction ():
    return
    {
        'patient_age':
        {
            'keyword':"age",'response':['15-25','25-40','40-50',''>50'],'serial':1

        },
        'patient_weight':
        {
            'keyword':"weight",'response':['40-50 kg','50-70 kg','70-90 kg',''>90 kg'],'serial':2

        },
        'patient_height':
        {
            'keyword':"height",'response':['4-5 ft','5-6 ft','6-7 ft'],'serial':1

        },
        'patient_gender':
        {
            'keyword':"gender",'response':['Male','Female','Unspecified'],'serial': 3
        }
}
```

Algorithm 1 Property extraction using keywords

The second category of data to be taken is keywords that are mapped to the parameters which help the system to check for the diagnosis process and enable the prediction strategy. Through these keywords, mapping and relationships are checked at a knowledge base where relevant diseases can be mapped and retrieved. There would be some constraints that are predefined and can take the value as yes/no that is stored in form of binary values 1 and 0 where detects the presence of disease or not respectively. framed for values entered for the parameters that are used to check for disease. The parameter considered are listed below along with the algorithm mentioning this tagging process.

- *Diabetes*: blood sugar level, weight, fasting sugar level, glucose tolerance, hemoglobin count.
- *Malaria*: temperature, plasmodium level, vomiting, Giemsa, sweat, muscle pain.

Similarly, the same process is carried out for different types of diseases as these are checked with disease ontology. The medical terms are already found to be stored in disease ontology that helps the processing of the information entered through web platform which is processed and knowledge is produced to the users.

Information processing using NLP

The data is collected under 3 different categories. Now the collected data is processed using different NLP packages. The content package available in python helps to extract all the data relevant to the keywords. Hence the extractors will do all types of preprocessing with the help of the NLP package. Information from the database is sent to the NLP module that does processing and does the extraction of relevant keywords. Just like that, the entire content cannot be processed hence the text is to be tokenized initially. The entire text is fragmented into lexical elements. The splitting process is carried out as the entire sentence is split into words. This process is carried out by a natural language toolkit (NLTK) tokenizer (Bird 2006). Now the words are taken and the relationship among the keyword and processed is checked using WordNet. This also addresses the words in form of nouns, verbs, adjectives, and adverbs.

The representation is mentioned as semantic space in a multidimensional way. Now this featured vector is given as the input to the classifier that returns the final disease label which is then transferred to the front end application and the additional feature can be sent through text message in their own native language.

Rule generation using the fuzzy logic system

In this paper, we focus on the fuzzy logic-based reasoning module (Nepal et al. 2005). It helps in collecting the data, preprocesses it, and gives the information back to the user in form of text messages both in web and mobile text. The entire process is categorized into different levels such as collection, process, symptoms extraction, and keyword mapping. Hence based on the keywords each disease is taken as a separate class and is associated with objects which are the symptoms of that disease. After this process, relevant fuzzy rules are generated to consider multiple diseases that are possible for the same type of symptoms. The fuzzy logic system using weighted calculations is used. The process is carried out based on the priority of weightage given is considered. The rule is based on the IF-THEN statement that is Mamdani fuzzy logic model. The rules go like this "If *x1* is *A1* and *y1* is *B1* then *z1* is *C1*" where *A1, B1, C1* are considered to be the fuzzy sets. For the sample we take the dataset to be represented as l, containing n features and *k* samples.

$$Ti = [\prec \mu FT1(a1), \mu FT2(a1)...\mu FTk(a1) \succ ... \prec \mu FT1(an), \mu FT2(an)...\mu FTk(an) \succ], \quad (1)$$

```
def data_diagnosis ():
    return
    {
        'patient_diabetes':
        {
            'keyword':"sugar level",'response':['70-80','80-100','100-120',''>120'],'serial':1

            'keyword':"GTL",'response':['112-124','125-136','137-145',''>145'],'serial':2

        },
        'patient_malaria':
        {
            'keyword':"fever",'response':['Yes, High (>103 F)','Yes, Mild (101-103 F)',
                      'Yes, Very Mild (99 - 101 F),no']

            'keyword':"Giemsa",'response':['Yes(123 Ml)','Mild (123-140 Ml)','High (>140 Ml)']
        }
}
```

Algorithm 2 Symptoms extraction using keywords

Disease prediction

The extracted words from all the previous steps are collected together which act as the base for the prediction of diseases. They are converted to feature vectors through the machine learning package supported by the python library.

where μFT represents the fuzzy term for the feature taken in which $F = ai$. In case the fuzzy variable is said to have k fuzzy terms then for the given each value the value for mapping and extraction is calculated as maximum among the computed value. After the computing of values, the overall process is been carried out in these steps. Data received are analyzed and extracted based on membership functions. Then using the fuzzy-based logic rule the mapping of keywords with parameters is done. After the mapping, the rules of weight calculation are done by fuzzy values. The rule attaining the highest score is taken as the final decision. Many classifiers are available which another vast area of research is. In our paper, we focus on the SVM classifier just to categorize the result received. We are training the classifier by the preprocessed information that is received through the NLP package further converted as feature vectors through word embedding.

Support vector machine for prediction

The SVM model (Claesen et al. 2014) represents different classes in a hyperplane. The SVM generates the hyperplane iteratively to reduce the error rate. The main objective of the SVM classifier is to partition the dataset into classes and identify the maximum marginal hyperplane. The SVM algorithm offers high accuracy and works well with high dimensional space. The Radial Basis Function kernel is mainly used for the SVM classification tasks and it maps the space into indefinite dimensional space. It is defined as follows:

$$k(a, a_j) = \exp(-\gamma + sum(a - a_j)^2) \quad (2)$$

The value of the γ is adjusted from 0 to 1 based on the algorithm and (a, a_j) is the input pair value.

Result in the native language

After the classifier process, the output result is been displayed in the specific login of the patient. Added to this feature the resulting observations either described by the doctor or through the prediction process of given input can be displaced in their own language. This process is implemented by using a python translation package. It takes in the semantics and other literals directly from Google and completes the translation process. Apart from English, it can include other languages and other country repositories such as youdao, baidu, iciba translation services. Finally, the dst parameter includes the specification of language preference. Since Tamil is the language of our preference we have set the dst value as "ta". If no value is set then it takes the value "zh-CN". It will result in the same language of input text. In case of exceptions when the proxy is not connected default caption is displayed by using the exception mechanism.

Results

The results of the project and their output are given in this section. The hyperparameters of the SVM are optimized using a Grid Search strategy (Belete et al. 2021). The parameters of the SVM optimized are namely gamma and the regularization parameters who were assigned the value 5.01E+10 and 14,201. For comparison, the classes are balanced and the vectors are normalized. The results were obtained by using 5-fold cross-validation and running the experiments a total of 20 times. Using the 5-fold cross-validation, the input data is partitioned into five groups. The dataset for this work is formed by gathering different clinical notes [Electronic Health Records (EHR)] from different medical organizations such as clinical labs, scan centers, hospitals, etc. A total of 4000 patient EHR was obtained from different specialists. The attributes present in the dataset which helps to identify different diseases such as Malaria, Blood Pressure, and Diabetes is provided in Table 1. After the NLP-based information processing is done in the dataset, the left out features are assigned as zero since no positive prediction can be made from them. The remaining data is partitioned into training and testing set in the ratio 8:2.

The login page of service providers and the patient view is given in Fig. 4. In case if the user has not registered then the sign-up process can be done which is represented in Fig. 5.

Once the user has logged in, they would be able to see their details as depicted in Figs. 6 and 7. It includes the profile of the user along with basic information, blood report details, other reports or tests taken, and final observations. Once the data is collected the fuzzy logic process is carried out in which the ruleset are generated and observations are made as shown in Fig. 8. Added to this if needed doctor mapping can be carried out. But while getting the information as a text message only the observation status alone is done.

The reports are viewable and can be downloaded at the patient end as shown in Fig. 9 so this makes the proposed system to be flexible that even though they migrate from one place to another the next set of treatments can be taken at any hospital with any doctor with just showing all these information and observations in the portal. Added to this feature the hospitals also will be able to keep track of their critical patients depicted in Fig. 10 so as to remind them of upcoming treatment dates. This is just an added feature as we are not focusing on the benefit of the hospital but instead only on the patient-user side.

Table 1 Features present in the dataset

Parameter	Range	Description
Age	24–98	–
Sex	Male/female	–
Body Mass Index (BMI)	20.2–40	People with a BMI greater than 35.14 are prone to diabetes or blood pressure.
Fasting plasma glucose (FPG)	100–125 mg/dL	Patients with FPG index Over 126 mg/dL are diabetic
Plasma glucose concentration (PPG)	70–110 mg/dL	Patients with PPG greater than 155 mg/dL is prone to diabetes
A hemoglobin A1c (HbA1c)	5.5–15.3%	It calculates the sugar level present in the red blood cell and the diabetic patients have an HbA1C value above 5.9%.
Plasmodium falciparum	Infected/Not	The plasmodium falciparum is mainly transmitted by the Anopheles mosquitoes. symptoms include headache, cough, tachycardia, tachypnea, chills, malaise, fatigue, diaphoresis (sweating), anorexia, nausea, vomiting, abdominal pain, diarrhea, arthralgias, and myalgias
Severe falciparum malaria	*P. falciparum* parasitemia > 10% (> 500,000/mcL).	It is acute malaria with confirmation of vital organ dysfunction including consciousness with a Glasgow Coma Score < 11, renal impairment, prostration, multiple convulsions, shock, pulmonary edema, jaundice, significant bleeding, severe malarial anemia, and hypoglycemia
Plasmodium malariae	Unknown fever > 38 °C mainly in the area with more malaria patients	The patients with *plasmodium malariae* may have more than one of these symptoms based on their lab results such as parasitemia, mild coagulopathy, anemia, thrombocytopenia, elevated transaminases, elevated blood urea nitrogen, and creatinine
A 2-h glucose tolerance test (2hPG)	140–200 mg/dL	A 2hPG value > or =200 mg/dl represents diabetes
Total cholesterol (TC)	89–277 mg/dL	A value ≥150 mg/dl indicates diabetes, blood pressure, or heart disease
Triglyceride level (TGL)	35–500 mg/dL	The TGL metric measures the level of fat in the blood and a value ≥ 150 mg/dl indicates diabetes, blood pressure, or heart disease
Cough	True/false	Symptom of COVID-19/malaria
Sore throat	True/false	Symptom of COVID-19
Shortness of breath	True/false	Symptom of COVID-19
Fever	True/false	Symptom of COVID-19/malaria
Headache	True/false	Symptom of COVID-19/malaria
Contact with a COVID-19 patient	Yes/no	A near contact with a COVID-19 patient can also sometimes confirm the presence of the disease

[a]PhI (1 mmol), $PhB(OH)_2$ (1 mmol), solvent (2 mL), and base (1.5 mmol)

Finally, the fuzzy logic module along with the SVM classifier is used for the final prediction mechanism. The results obtained are shown in Fig. 11.

The prediction is sent to the user in the English language as default. When the option of text preference is selected then information reaches in their preferred language. In our research, we have focused on and considered one language Tamil. The text of prediction results reached in mobile is displayed in Figs. 12 and 13 given below.

Performance metrics

The proposed methodology is evaluated using different performance metrics namely accuracy (A), Error Rate, Precision (P), Recall (R), and F1-Score. If a disease is correctly categorized according to its disease class, it is represented as a true positive (T_{pos}), whereas if the patient is not affected by any disease, it should output normal, which is known as a true negative (T_{neg}). If a disease is erroneously classified, it is referred to as a false positive (F_{pos}), and if a healthy person is identified as having a certain disease, it is referred to as a false negative (F_{neg}).

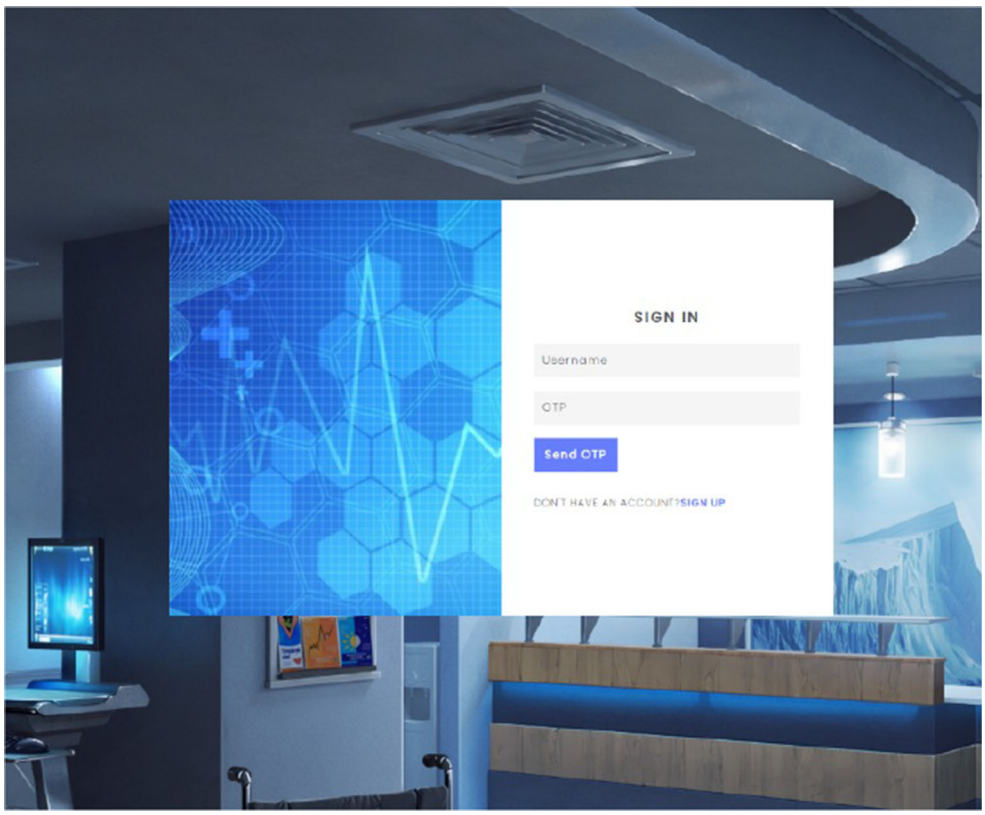

Fig. 4 Login page

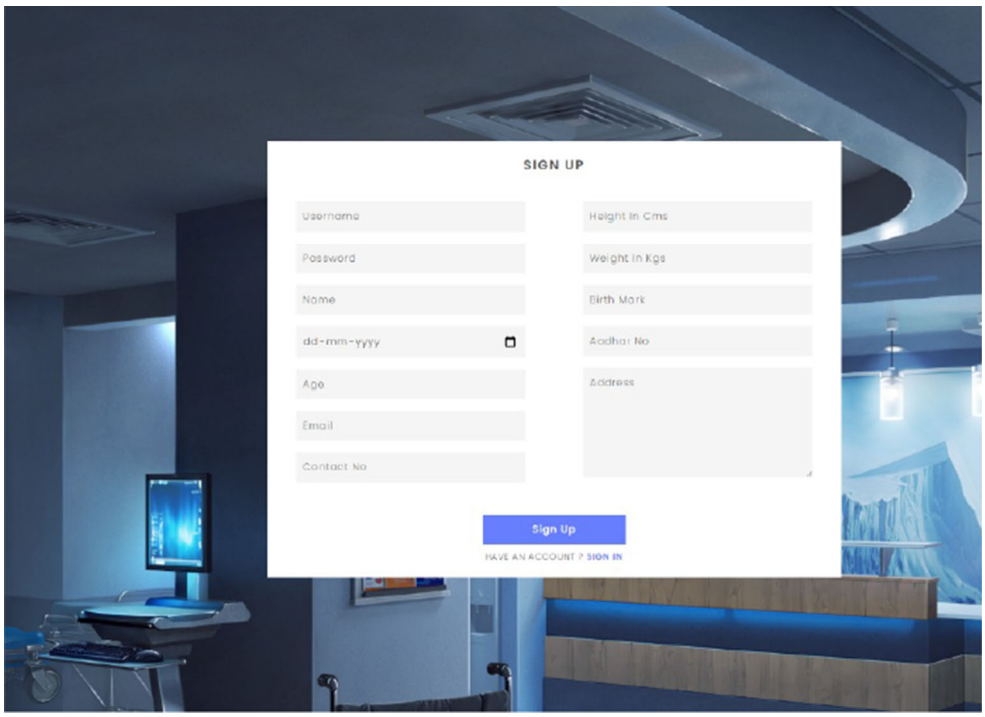

Fig. 5 Sign up the process of user

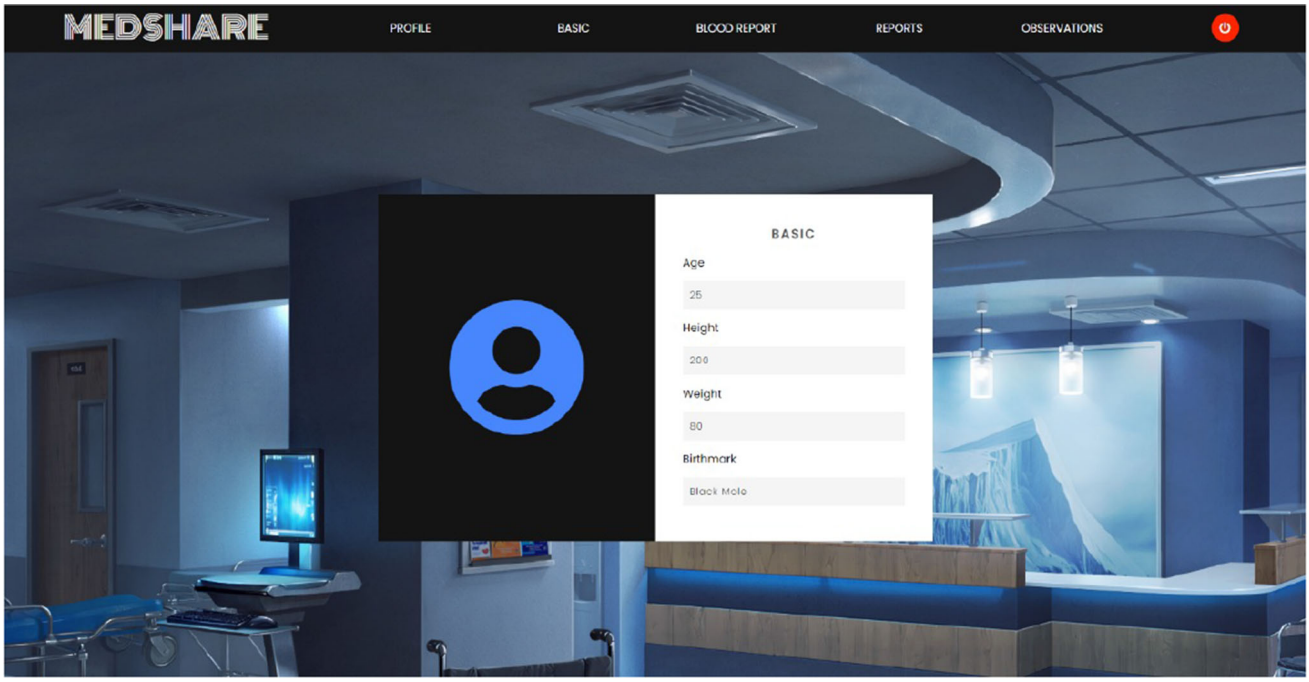

Fig. 6 Data collection view

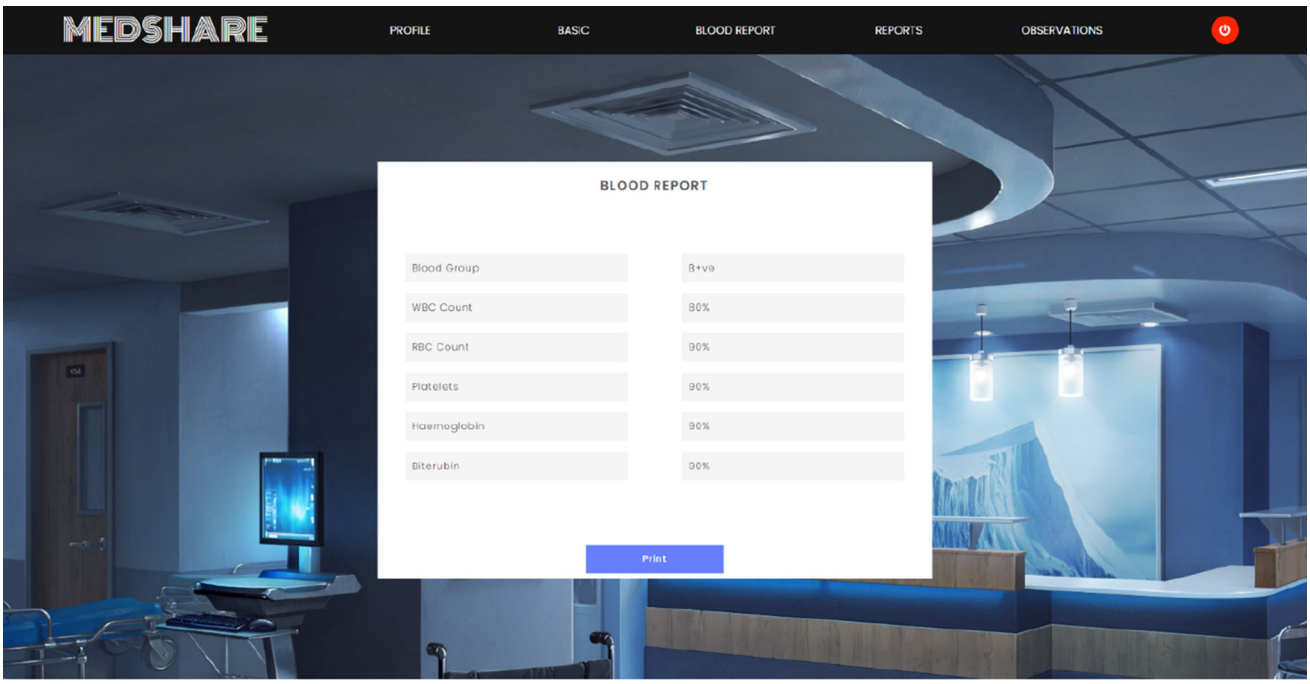

Fig. 7 Blood test parameters collection

The accuracy mainly measures the ability of a machine learning classifier to classify the input disease correctly from the blood test report. The error rate, which is the inverse of accuracy, computes the number of inaccurate predictions made by the model. Precision is defined as the proportion of correctly identified positive disease predictions among all predicted positive instances. The recall is defined as the proportion of correctly diagnosed disease predictions to positive predictions made. The F1-Score can indicate the tradeoff between precision and recall, as well as the imbalanced class distribution.

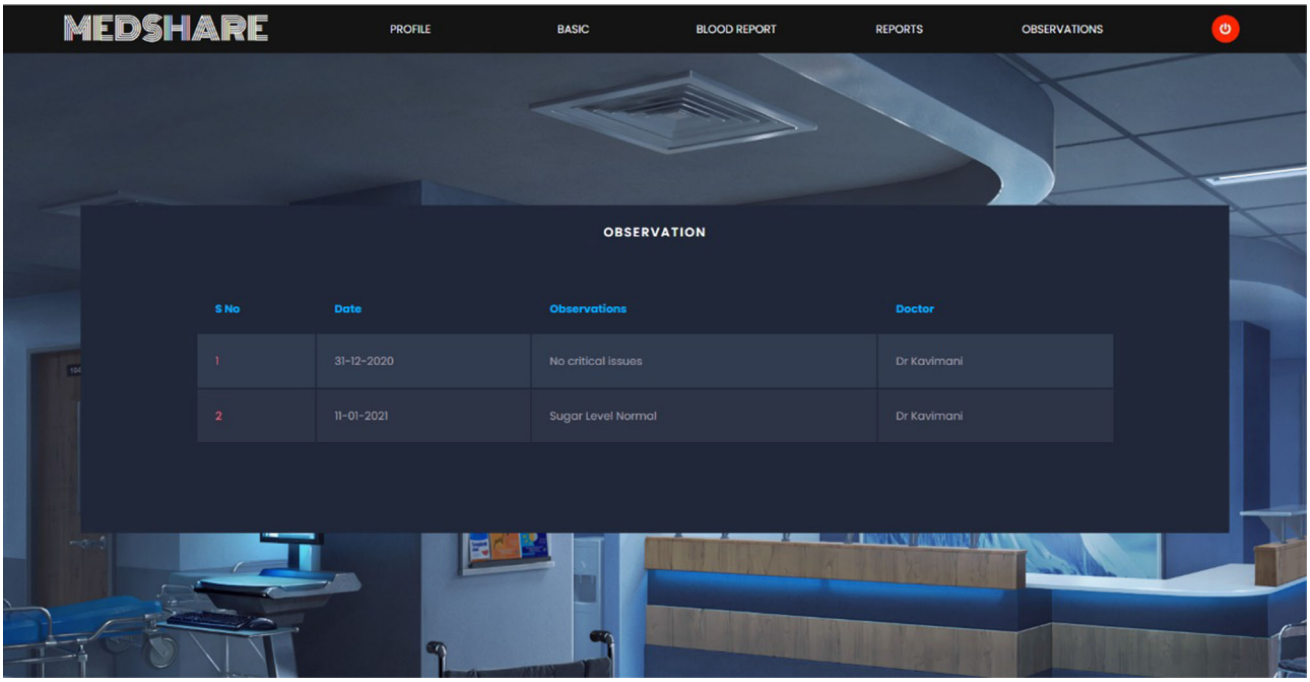

Fig. 8 Extraction of keywords

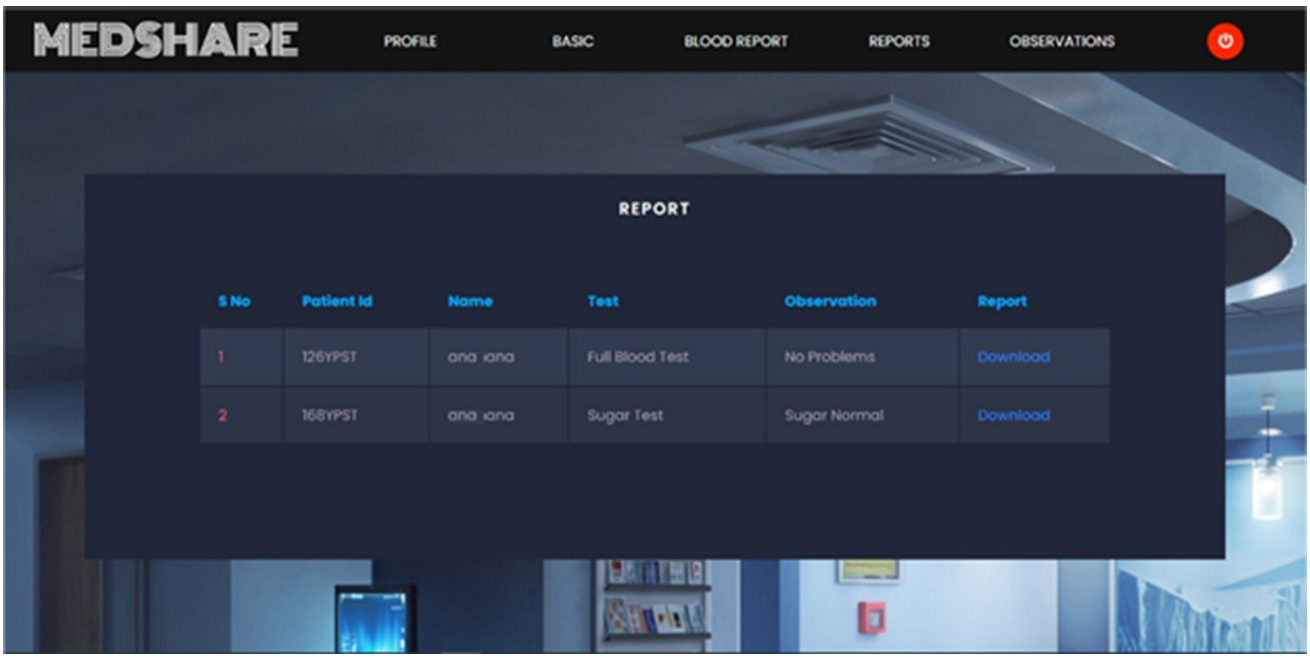

Fig. 9 Reports view

$$Accuracy = \frac{T_{POS} + T_{NEG}}{T_{POS} + T_{NEG} + F_{POS} + F_{NEG}} \quad (3)$$

$$Error\_Rate = 1 - A \quad (4)$$

$$P = \frac{T_{POS}}{T_{POS} + F_{POS}} \quad (5)$$

$$R = \frac{T_{POS}}{T_{POS} + F_{NEG}} \quad (6)$$

$$F1 - score = 2 * \frac{P * R}{P + R} \quad (7)$$

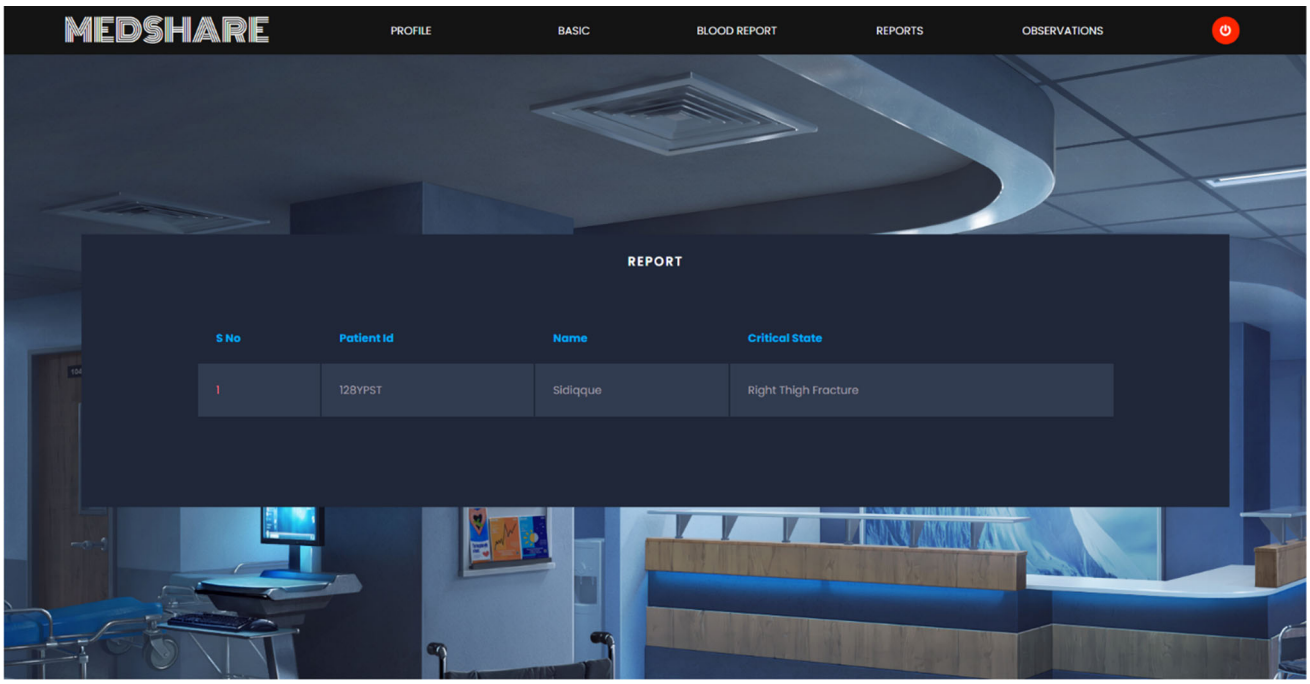

Fig. 10 Critical status of patients

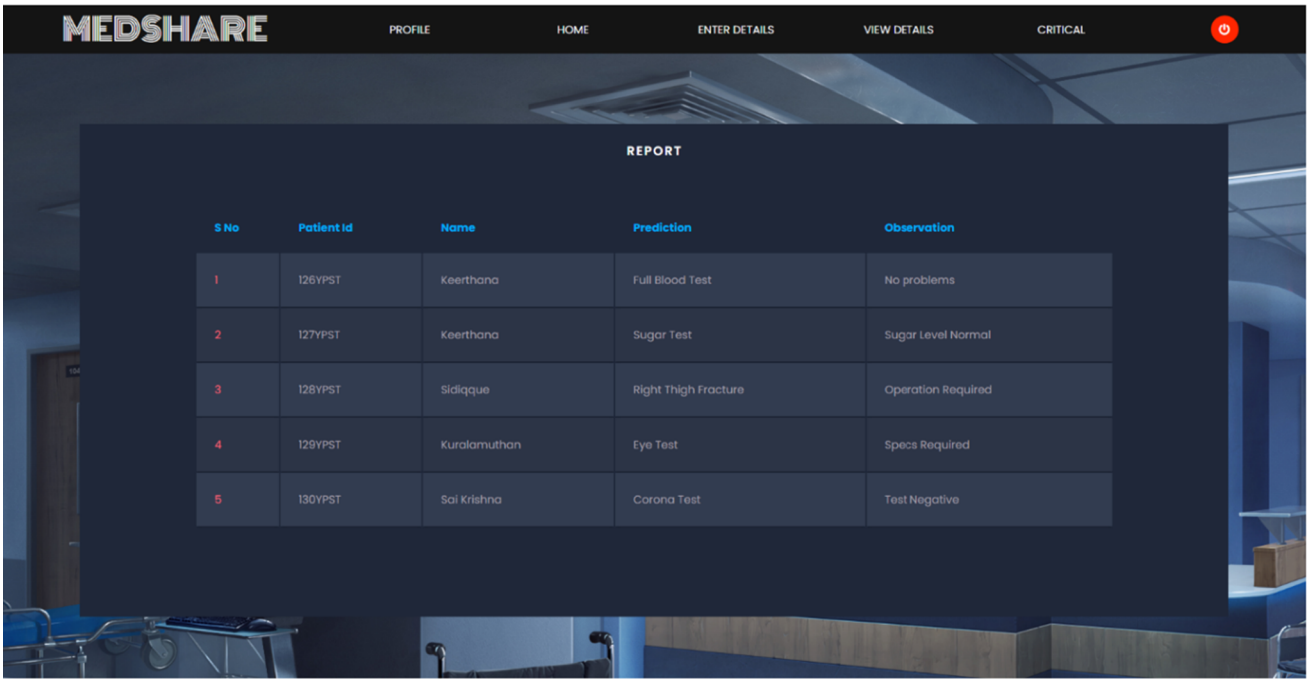

Fig. 11 Prediction based on data integration

Performance evaluation

The comparison results are provided in this section. Initially, the proposed methodology is evaluated in terms of accuracy by comparing it with other techniques such as decision tree (Johnson et al. 2002) and Naïve Bayes classifier (Eyheramendy et al. 2003). The SVM classifier used here results in 89% prediction accuracy in Fig. 14. The same is compared with the decision tree classifier and Naïve Bayes which shows that SVM to be the best. In further process, multiple comparisons or ensemble method

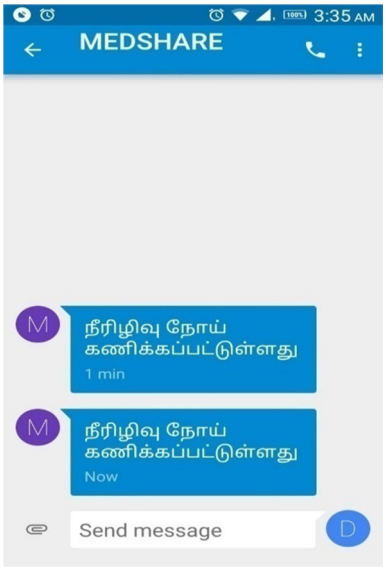

Fig. 12 Prediction result in Tamil

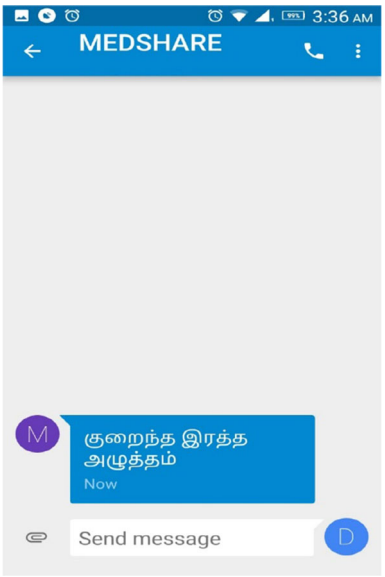

Fig. 13 Prediction result in Tamil

can be carried out which gives efficiency more than what have attained.

The proposed methodology is compared in terms of precision, recall, and F1-Measure with other techniques such as Decision tree and Naïve Bayes classifier. The proposed methodology offers a precision, recall, and F1-score of 92%, 91%, and 91% when compared to the decision tree and naïve Bayes techniques (Fig. 15). The results obtained for the Naïve Bayes and Decision tree classifier is presented in Table 1. The error rate results are presented in Fig. 16 and the proposed methodology also provides a lower error rate when compared with others.

Table 2 provides the comparative analysis of the proposed methodology with the Naïve Bayes and Decision tree technique in terms of Diabetes, Blood Pressure, Malaria, and COVID19 prediction. The results are verified for the predictions in Tamil and the vectors were trained using an English corpus. From the results shown in Table 3, we can see that the proposed methodology obtains the best results when compared to the other techniques. The results of the decision tree and Naïve Bayes are quite similar in value but the Decision tree offers the lower F1-score. The final results indicate that the proposed technique acts as an outstanding disease prediction tool for users and clinicians in their native language.

Usability testing

The usability of the proposed medical application is tested using the chatbot system usability scale present in the literature (Cameron et al. 2018). The usability is mainly tested in terms of effectiveness, ease of use, and efficiency. We have designed only eight questions and the answer value falls in the range of 0 to 100. The different questions (A1′–A8) designed are presented below:

A1: The application is easy to use.
A2: I am willing to use this application for my future appointments.
A3: The different functionalities are very well integrated.
A4: It is easy to be used by laypeople.
A5: The MEDSHARE application is very complex to use.
A6: This application needs a little bit of practice to be used.
A7: The GUI design is a little bit complex.
A8: Too much advertisement and needs frequent updates.

There was a total of 30 participants among which 17 were male and 13 were female from an age gap of 13–59. The study lasted a total of 40 min. The scores were

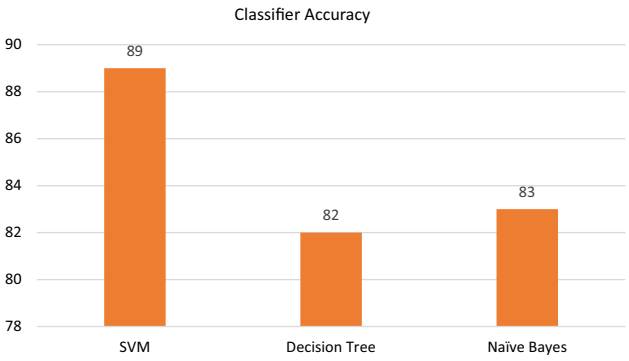

Fig. 14 Prediction accuracy

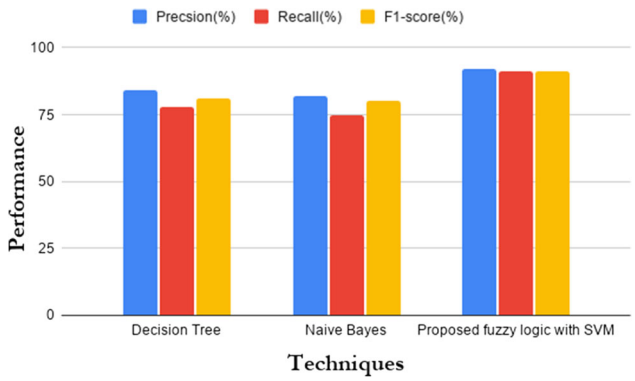

Fig. 15 Comparison in terms of precision, recall, and F1-score

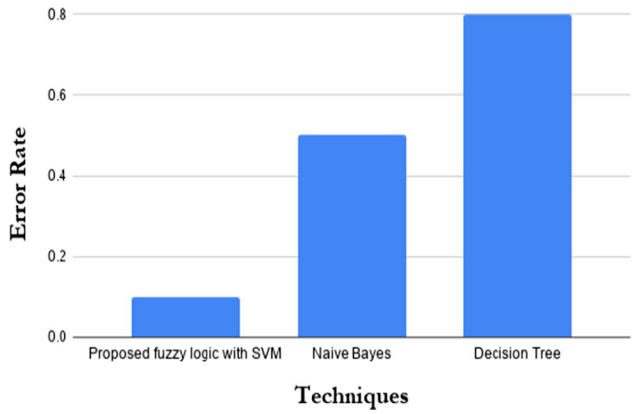

Fig. 16 Performance evaluation using error rate

calculated by subtracting one if questions A1, A2, A3, and A4 were selected, and subtracting five if questions A5–A8 were selected. The results are then multiplied by 2.5 to yield the final evaluation score. The scores above 80 were considered excellent, a score above 70 was considered good, and scores above 60 were considered okay and scores below 50 were considered bad. The results are shown in Fig. 17, and the values obtained for the questions A1–A4 are very high which shows that the proposed application is easy to use, can be understood by laypeople, and the functionalities are well integrated. A higher value was obtained for question A4 which says that the proposed approach is easily understandable even by a layperson. A6 is the question that had a higher score regarding the complexity of the MEDSHARE application. Half of the candidates said that they need a little bit more practice to operate this application.

Discussion

The healthcare records contain structure, unstructured data, and also other reports such as blood tests, sugar tests, etc. Hence the queries are collected from the users and information is retrieved from the available data and returned to the user in a structured manner. Some of the techniques added to this in NLP helps in providing additional support such as diagnosis. But combining all these features is still a challenging process as data integration needs to be carried out. The techniques can process a large amount of text data which can be converted to knowledge to access by the patients which are at times self-understanding. Eventhough the NLP systems for healthcare applications achieved multiple benefits, it often struggles in processing the heterogeneous data obtained from various medical institutions due to the presence of different languages. The symptoms present in different sublanguages often hinder performance. Based on our experimental outcomes, we can confirm that the usage of WordNet, UMLS, and Disease

Table 2 Comparative analysis results

Techniques	Precsion (%)	Recall (%)	F1-score (%)
Decision tree	84	78	81
Naive Bayes	82	75	80
Proposed fuzzy logic with SVM	92	91	91

Table 3 Comparative results obtained for different diseases using F1-score

Techniques	F1-score(%)							
	Diabetes		Blood Pressure		Malaria		COVID 19	
	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
Decision tree	0.76	0.78	0.76	0.75	0.77	0.78	0.78	0.79
Naive Bayes	0.81	0.83	0.80	0.82	0.81	0.82	0.83	0.82
Proposed fuzzy logic with SVM	0.91	0.95	0.93	0.96	0.95	0.93	0.95	0.96

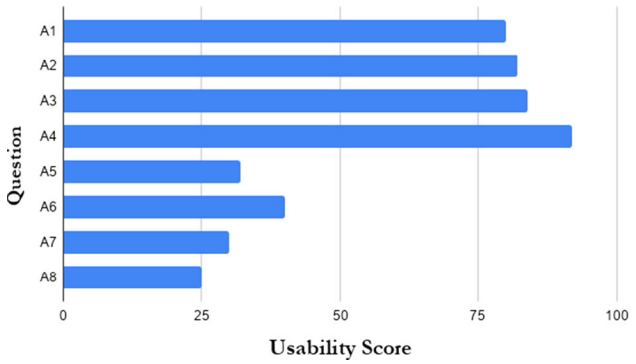

Fig. 17 Usability testing

ontology has improved the information extraction process and increased the accuracy of the MEDSHARE application. Our proposed methodology offers optimal performance even when evaluated with noisy medical datasets.

The performance of the proposed methodology is even higher than the state-of-art techniques such as SYNODOS NLP solution (Tvardik et al. 2018); AltibbiVec (Habib et al. 2021), Deep learning (Shickel et al. et al. 2017), Contextual Chatbot (Kandpal et al. (2020), and NLP (Jungmann et al. 2020). The contextual chatbot application is designed in the English language which is often complex to be used by laypeople. The performance of the automatic phenotype extraction approach proposed by Liu et al. (2019) often declined due to the size of the training sample used. To increase the performance of the classifier the size of the training samples needs to be increased. This problem exists for different state-of-art approaches proposed. The NLP technique proposed by Solomon et al. (2021) to identify the aortic stenosis patients often struggled with the patient-specific concept recognition task due to the missed follow-up visits and membership cancellation of the patients. The AltibbiVec proposed by Habib et al. (2021) is only applicable to large datasets to yield effective contextualized word representation outcomes. Based on these results we can observe the use of the NLP and Machine learning techniques is efficient for healthcare applications.

With the help of these techniques, we can easily diagnose complex diseases such as malaria, diabetes, blood pressure, etc. The NLP packages used in our proposed methodology efficiently extract a large number of acronyms with varying modes of writing for effective disease diagnosis. The SVM classifier accurately predicts the disease of the patient. The fuzzy rule generation framework generates the appropriate rules to identify the different diseases with the same symptoms. The symptoms of malaria and COVID19 can sometimes collide and the same can be also applicable to diabetes and blood pressure. The proposed methodology has some limitations which are delineated below. The proposed work doesn't extract the patient's phenotypes which is valuable when it comes to patient-specific concept recognition tasks. In the future, we plan to overcome this issue using a majority voting ensemble technique. In some cases, a novel dataset with a low number of features is prone to the overfitting and curse of dimensionality problem which can also hinder real-time performance.

Conclusions

The access of information to healthcare records must be done promptly to avoid dreadful situations. Hence this application MEDSHARE makes users access their own medical information wherever or whenever necessary. In addition, the application has a feature that automatically combines the diagnosis process with the data stored in the application. Thus fuzzy logic module generates the ruleset that enables embedding technique with NLP. Then using the SVM classifier the prediction process is said to have 89% of accuracy and seems to be proven highest when compared to the decision tree and Naïve Bayes classifier. To make this application to be enhanced and for further betterment, additional features can be added such as mapping of doctors in case of emergencies, alert notifications when exceeding the threshold, and multiple language translation. In the future can be optimized by including ensemble methods for prediction purposes.

Availability of data and material Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Human and animal rights This article does not contain any studies with human or animal subjects performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

Ash JS, Berg M, Enrico Coiera (2004) Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. J Am Med Inform Assoc 11(2):104–112

Belete DM, Huchaiah MD (2021) Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results.Int J Comput Appl

Bird S (2006) NLTK: the natural language toolkit. In: Proceedings of the COLING/ACL 2006 interactive presentation sessions, pp 69–72

Bodenreider O (2004) The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 32(-suppl_1):D267–D270

Cameron G, Cameron D, Megaw G, Bond R, Mulvenna M, O'Neill S, Armour C, McTear M (2018) Assessing the usability of a chatbot for mental health care. In: International conference on internet science. Springer, Cham, pp 121–132

Claesen M, De Smet F, Johan AK, Suykens, De Moor B (2014). Fast prediction with SVM models containing RBF kernels. arXiv preprint arXiv:1403.0736

Corral-Plaza D, Medina-Bulo I, Ortiz G, Boubeta-Puig J, UCASE Software Engineering Research Group (2020) A stream processing architecture for heterogeneous data sources in the internet of things. Comput Stand Interfaces 70:103426

Coulter A, Collins A (2011) Making shared decision-making a reality. King's Fund, London

Dick RS, Elaine B, Steen, Detmer DE (eds) (1997) eds. The computer-based patient record: an essential technology for health care. National Academies Press

Eyheramendy S, Lewis DD, Madigan D (2003) On the naive bayes model for text categorization. In: International workshop on artificial intelligence and statistics. PMLR, pp. 93-100

Ghasemzadeh H, Ostadabbas S, Pantelopoulos A (2012) Wireless medical-embedded systems: a review of signal-processing techniques for classification. IEEE Sens J 13(2):423–437

Gowthul Alam MM, Baulkani S (2019b) Local and global characteristics-based kernel hybridization to increase optimal support vector machine performance for stock market prediction. Knowl Inf Syst 60(2):971–1000

Gowthul Alam MM, Baulkani S (2017) Reformulated query-based document retrieval using optimised kernel fuzzy clustering algorithm. Int J Bus Intell Data Min 12(3):299

Gowthul Alam MM, Baulkani S (2019a) Geometric structure information based multi-objective function to increase fuzzy clustering performance with artificial and real-life data. Soft Comput 23(4):1079–1098

Habib M, Faris M, Alomari A, Faris H (2021) AltibbiVec: a word embedding model for medical and health applications in Arabic language. IEEE Access

Hardin M, Chhieng DC (2007) Data mining and clinical decision support systems. In: Clinical decision support systems. Springer, New York, NY, pp 44–63

Haseena KS, Anees S, Madheswari N (2014) Power optimization using EPAR protocol in MANET. Int J Innov Sci Eng Technol 6:430–436

Johnson DE, Oles FJ, Zhang T, Goetz T (2002) A decision-tree-based symbolic rule induction system for text categorization. IBM Syst J 41(3):428–437

Jungmann F, Kämpgen B, Mildenberger P, Tsaur I, Jorg T, Düber C, Mildenberger P, Kloeckner R (2020) Towards data-driven medical imaging using natural language processing in patients with suspected urolithiasis. Int J Med Inf 137:104106

Kandpal P, Jasnani K, Raut R, Bhorge S (2020) Contextual chatbot for healthcare purposes (using deep learning). In: 2020 fourth world conference on smart trends in systems, security and sustainability (WorldS4). IEEE, pp 625–634

Kaur B (2020) Disasters and exemplified vulnerabilities in a cramped Public Health Infrastructure in India. Int J Disaster Risk Manag 2(1):15–22

Kaur H, Sohn S, Wi C-I, Ryu E, Park MA, Bachman K, Kita H et al (2018) Automated chart review utilizing natural language processing algorithm for asthma predictive index. BMC Pulm Med 18(1):1–9

Kavitha D, Ravikumar S (2021) IOT and context-aware learning-based optimal neural network model for real-time health monitoring. Trans Emerg Telecommun Technolog 32(1):e4132

Kothmayr T, Schmitt C, Wen Hu, Brünig M, Carle G (2013) DTLS based security and two-way authentication for the Internet of Things. Ad Hoc Netw 11(8):2710–2723

Légaré F, Witteman HO (2013) Shared decision making: examining key elements and barriers to adoption into routine clinical practice. Health Aff 32(2):276–284

Li X, Wang H, He H, Du J, Chen J, Wu J (2019) Intelligent diagnosis with Chinese electronic medical records based on convolutional neural networks. BMC Bioinform 20(1):1–12

Liu C, Ta CN, Rogers JR, Li Z, Lee J, Butler AM, Shang N, Kury FSP, Wang L, Shen F, Liu H (2019) Ensembles of natural language processing systems for portable phenotyping solutions. J Biomed Inf 100:103318

Morato J, Marzal MA, Lloréns J, Moreiro J (2004) Wordnet applications. In: Proceedings of GWC, pp 20–23

Nanjappan M, Albert P (2019) Hybrid-based novel approach for resource scheduling using MCFCM and PSO in cloud computing environment. Concurr Comput Practice Exp. https://doi.org/10.1002/cpe.5517

Nanjappan M, Natesan G, Krishnadoss P (2021) An adaptive neuro-fuzzy inference system and black widow optimization approach for optimalresource utilization and task scheduling in a cloud environment. Wireless Pers Comm 121(3):1891–1916. https://doi.org/10.1007/s11277-021-08744-1

Nepal B, Monplaisir L, Singh N (2005) Integrated fuzzy logic-based model for product modularization during concept development phase. Int J Prod Econ 96(2):157–174

Névéol A, Zweigenbaum P (2015) Clinical natural language processing in 2014: foundational methods supporting efficient healthcare. Yearb Med Inform 10(1):194

Popowich F (2005) Using text mining and natural language processing for health care claims processing. ACM SIGKDD Explor Newsl 7(1):59–66

Ravikumar S, Kavitha D (2021) CNN-OHGS: CNN-oppositional-based Henry gas solubility optimization model for autonomous vehicle control system. J Field Robot

Ravikumar S, Kavitha D (2020) IoT based home monitoring system with secure data storage by Keccak–Chaotic sequence in cloud server. J Ambient Intell Human Comput 1–13

Rejeesh MR, Thejaswini PMOTF (2020) Multi-objective optimal trilateral filtering based partial moving frame algorithm for image denoising. Multimed Tools Appl 79:28411–28430. https://doi.org/10.1007/s11042-020-09234-5

Roberts A (2017) Language, structure, and reuse in the electronic health record. AMA J Ethics19(3):281–288

Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, Feng G, Kibbe WA (2012) Disease ontology: a backbone for disease semantic integration. Nucleic Acids Res 40(D1):D940–D946

Shickel B, Tighe PJ, Bihorac A, Rashidi P (2017) Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 22(5):1589–1604

Solomon MD, Tabada G, Allen A, Sung SH, Go AS (2021) Large-scale identification of aortic stenosis and its severity using natural language processing on electronic health records.Cardiovasc Digital Health J

Srivastava P, Singh N (2020) Automatized medical Chatbot (Medibot). In: 2020 international conference on power electronics & IoT applications in renewable energy and its control (PARC). IEEE, pp 351–354

Sivaranjani J, Madheswari AN (2017, March). A novel technique of motif discovery for medical big data using hadoop. In: 2017

conference on emerging devices and smart systems (ICEDSS). IEEE, pp 214–217

Srinivasan A, Madheswari AN (2018) The role of smart personal assistant for improving personal healthcare. International Journal of Advanced Engineering, Management and Science, 4(11), p.268274

Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Jun'ichi Tsujii (2012) BRAT: a web-based tool for NLP-assisted text annotation. In: Proceedings of the demonstrations at the 13th conference of the European chapter of the Association for Computational Linguistics, pp 102–107

Sundararaj V, Selvi M (2021) Opposition grasshopper optimizer based multimedia data distribution using user evaluation strategy. Multimed Tools Appl 80(19):29875–29891

Sundararaj V (2019) Optimised denoising scheme via opposition-based self-adaptive learning PSO algorithm for wavelet-based ECG signal noise reduction. Int J BioMed Eng Technol 31(4):325

Sundararaj V, Anoop V, Dixit P, Arjaria A, Chourasia U, Bhambri P, Rejeesh MR, Sundararaj R (2020) CCGPA-MPPT: cauchy preferential crossover-based global pollination algorithm for MPPT in photovoltaic system. Prog Photovolt Res Appl 28(11):1128–1145

Tvardik N, Kergourlay I, Bittar A, Segond F, Darmoni S, Metzger MH (2018) Accuracy of using natural language processing methods for identifying healthcare-associated infections. Int J Med Inf 117:96–102

Sundararaj V (2016) An efficient threshold prediction scheme for wavelet based ECG signal noise reduction using variable step size firefly algorithm. Int J Intell Eng Syst 9(3):117–126

Walczak S (2005) Artificial neural network medical decision support tool: predicting transfusion requirements of ER patients. IEEE Trans Inf Technol Biomed 9(3):468–474

Walshe K, Rundall TG (2001) Evidence-based management: from theory to practice in health care. Milbank Q 79(3):429–457

Wu J, Chen J, Zhang Q, Tang Z (1990) Transformation and identification of recombinant plasmid pAT153 containing HCMV gene Hind III F fragment and its clinical application. J West China Univ Med Sci 21(2):117–120

Xie Q, Sundararaj V, Mr R (2021) Analyzing the factors affecting the attitude of public toward lockdown, institutional trust, and civic engagement activities. J Commun Psychol

Yu B, He Z, Xing A, Lustria MLA (2020) An informatics framework to assess consumer health language complexity differences: proof-of-concept study. J Med Internet Res 22(5):e16795

Zhang Y, Fong S, Fiaidhi J, Mohammed S (2012) Real-time clinical decision support system with data stream mining. J Biomed Biotechnol

Zhong Q-Y, Mittal LP, Nathan MD, Brown KM, González DK, Cai T, Finan S et al (2019) Use of natural language processing in electronic medical records to identify pregnant women with suicidal behavior: towards a solution to the complex classification problem. Eur J Epidemiol 34(2):153–162