International Journal of Advanced Technology and Engineering Exploration, Vol 9(90)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2021.875069

Research Article

ETL for disease indicators using brute force rule-based NLP algorithm and metadata exploration

Ifra Altaf[1], Muheet Ahmed Butt[2*] and Majid Zaman[3]
Research Scholar, Department of Computer Sciences, University of Kashmir, Srinagar, J&K, India[1]
Scientist D, Department of Computer Sciences, University of Kashmir, Srinagar, J&K, India[2]
Scientist E, Directorate of IT&SS, University of Kashmir, Srinagar, J&K, India[3]

Received: 21-October-2021; Revised: 10-May-2022; Accepted: 14-May-2022

Abstract
As data driven decisions are based on facts, data collection can be used to lay a foundation for decision-making irrespective of industry. With the decision-making capability provided by the data from various digital medical records, the doctors can provide a precise diagnosis and a sufficient treatment by fitting together fundamentally different disease symptoms. This data manuscript describes the preparation procedure of a diabetes dataset from the panels of liver and lipid profile. The data is collected from a medical center in Srinagar, Jammu and Kashmir in the form of unstructured data reports. The unstructured data is extracted on the basis of the metadata of the source document; the required data field values of different tests are extracted from the intermediate file using the brute force pattern matching heuristics and integrated together to fill the relational database. The database can be used for further descriptive, exploratory as well as predictive data analysis and can be helpful in diagnosing and predicting the diabetes disease of the liver and lipid panels. This paper presents a novel concept to predict and detect one disease from the markers of other related disease/s as a way to fill the theoretical research gap. The detection rate achieved by our proposed brute force rule-based natural language processing (NLP) algorithm is recorded as 98.44%.

Keywords
PDF scraping, Unstructured data, Diagnostic lab reports, Heuristics, Brute force, Natural language processing, Metadata, Information extraction.

1.Introduction
Healthcare data is generated at an extensive rate every single second making the healthcare data growth an exponential problem. Every type of data related to the health status of an individual is termed as health data that comprises of the clinical metrics together with behavioural, ecological, and economic and social [1] health related information. Data driven health technologies [2−6] are undoubtedly bringing the ground-breaking changes to revitalize the healthcare sector. They can even arrange for assistance in diagnosing and predicting certain diseases with the help of using the diagnostic laboratory test reports [7−10] to deliver the best results. Likewise, this can lead to improved effectiveness and lower the expenditures of carrying out the biochemical tests. The advance in operational proficiency can lead to an improved care.

The several sources of health data consist of hospital records, medical histories of patients, administrative data, results of medical check-ups and examinations, [11] etc. Healthcare data also take account of a massive amount of medical data in the form of electronic medical record (EMR) and electronic health record (EHR). EMR is the subset of EHR [12] and are time and again used interchangeably. For diagnosis and treatment, mostly EMR is used. EHR is digital records of health data that contain a lot beyond EMR. Lab tests and diagnostic procedures are included in the EMRs. A diagnostic report is a set of biochemical tests that provide information regarding a person's health. The collected data having diagnostic labels can benefit in the automatic diagnosis of various diseases. Consequently, it can arrange for a prompt diagnostic reference for doctors other than helping to predict the disease ahead of time.

*Author for correspondence

The high availability of data makes machine learning and analytics an imperative technique for demonstrating the human process in the medical field. There are many chronic diseases and according to the World Health Organization, diabetes is one of them [13].

Diabetes is a health condition which doesn't allow glucose to move into the body cells in order to produce energy. An improved treatment can be assured if this disease is predicted at an early stage. Unbalanced liver enzymes in diabetic people act as the hepatocellular damage indicators [14]. Diabetes lowers "good" cholesterol levels and raises triglyceride levels [15]. The liver functioning is also associated with lipid metabolism. Since diabetes is associated with liver disorders and abnormal lipid levels, therefore the liver and lipid abnormalities can become significant attributes for the early prediction of diabetes disease. The relationship among the differing attributes of the liver and lipid panel tests can also be characterized.

Healthcare data is very inimitable and challenging to measure because the data is located in multiple places and also the data is mostly present in unstructured format rather than the structured one. The diagnostic laboratory tests generally use the fully formatted report such as a portable document format (PDF) because of their handy and human readable property. However, this format does not have a standard layout and is unstructured [16]. The diagnostic test reports are not kept in one single PDF file, but may possibly produce multiple PDF files. To access the information stored in the PDFs for data analysis, hence becomes a complicated task. With the unstructured data one cannot easily automate the data entry process and therefore the subsequent process like filling of the records cannot be automated. It is very demanding to work using text in PDF files since they are acknowledged as a picture [17]. Also the problem with the PDF files is that they can be unevenly outsized and cannot be expected to be the means of permanent storage. It is time consuming to manually examine the data stored in these unstructured arrangements.

The structured data on the other hand has a dedicated framework that facilitates analysis. It increases accessibility to the data because the data can be stored in one place and is responsible for effective data integration. The structured set of data in the form of database can make available a solution to the problems faced by the unstructured medical diagnostic laboratory PDF reports. The quality of the information also increases [18] with the better data management, which in turn helps in faster and improved decision making. A database can be easily accessed and managed by means of the programming language, structured query language (SQL).

Unstructured data is also known as the dark data. A lot of time is involved in the manual extraction of the unstructured data from PDF files. So, it becomes necessary to convert the unstructured data into the structured one that could be used for the data exploration. An inordinate drawback of present PDF scrapping approaches is their essential dependence to the application area. A relevant branch of natural language processing (NLP) that is concerned with the extraction of structured data from unstructured data is the information extraction (IE) which is grounded on the predefined information in the form of rules. Rule based approaches [19] use different rules to extract the content of documents and perform necessary actions. The central point of this paper is to present a PDF scrapping method based upon the metadata analysis. Since the PDFs lack the generalization, so it becomes very difficult to differentiate the different parts of the PDF. The metadata of the PDF can help to overcome this problem. We developed a brute force rule-based NLP approach with metadata exploration technique in order to extract the data from the PDF files by creating a generalized structure for every PDF. The method converts the unstructured data into the structured one for further processing and helps in gaining valuable insights from the data. Also, the paper presents certain valuable distribution information of the paper as well as visualizations acquired from the extracted dataset. As per our knowledge, this is the first study to use a brute force rule-based NLP approach with metadata exploration to extract information from clinical diagnostics laboratory reports.

1.1Motivation

The biggest motivation behind going for this research work with a novel idea of predicting one disease from the markers of another disease is the

- Availability of data in the form of unstructured PDF reports.
- Connection between the markers of certain diseases.

According to our observation, the most common laboratory test that the patients undergo is the blood sugar test accompanied habitually by the lipid profile panel (LPP) and liver function test (LFT) tests. Also,

some of the diseases have interrelated symptomatic attributes; therefore the contingent diseases can be diagnosed while treating another disease. This can also help in predicting the onset of other disease at an early stage.

1.2Objectives

A lot of data can be generated from these unstructured clinical PDF reports which can be put to valuable use. The derived patterns from the data can assist the medical experts with accurate and enhanced diagnosis. The main objectives of this research study are:

- Extraction of data from the unstructured clinical PDF reports.
- Integration and formation of a structured database from the unstructured extracted data.
- Classification of data-records into the appropriate target classes.
- Outsets of the extracted dataset.

The objectives are attained by extracting the information about PDFs using the brute force rule-based NLP approach. The PDF structure is parsed first and based upon the metadata information it gets divided into the different sections. These sections are analysed individually to identify the common patterns based upon the rules that are formed according to the values that we want to extract from the reports.

1.3Contribution

Most of the existing methods for diabetes diagnosis and early prediction only take advantage of either the standard datasets or the structured data stored in the EMRs. The contributions of our research study are as under:

- Improved rule based NLP extraction method based on metadata analysis.
- A new disease dataset that promotes the novel idea of predicting diabetes disease from the markers of another disease/s.
- Exploration of available laboratory test indicators of the diabetes disease on top of the traditional markers.
- Approach to extract the significant disease attribute values from PDF into a structured form.
- Presents a good observation system where information about the occurrence of other connected diseases can be provided beforehand.

The rest of the paper is organized in the following sections. Section 2 discusses the literature review. Section 3 provides the methodology. Section 4 discusses the results. Discussion of results is presented in section 5 and lastly, Section 6 concludes the paper and states the future work.

2.Literature review

The researchers have proposed different rule-based techniques from the past several years for PDF extraction techniques. We provide a brief summary of our analysis of the contemporary approaches, particularly focusing on the rule-based approach.

Ahmad et al. (2016) [20] suggested a rule-based approach where the metadata of PDF documents was extracted using its textual and physical features. The approach was assessed on a very small dataset. The output generated influenced the extraction of the metadata. The F-score of the approach was recorded as 0.77.

Sateli and Witte (2016) [21] proposed a hybrid approach wherein they combined the rule-based approach with the machine learning approaches. They used the supervised classifiers and pattern, structure and style based features to extract the PDF data. The limitation that the approach faced was related to the small size of the dataset. The F-score of the approach was 0.63.

Klampfl and Kern (2016) [22] suggested a machine learning based approach on PDF documents to extract metadata. Their unsupervised approach combined rule-based approach with the named entity recognition. A logical structure analysis of the PDF document is followed by IE using the unsupervised learning. The F-score of the approach was recorded as 0.59. The limitation of the proposed approach was linked to the very small size of the training dataset of PDF documents.

Azimjonov and Alikhanov (2018) [23] recommended a rule-based approach to extract metadata of the PDF documents in the form of extensible mark-up language (XML). For metadata extraction, the approach could not have similar results for all the datasets.

Hashmi et al. (2020) [19] analyzed that since PDFs have a complex logical structure than other formats, therefore the proposed extraction systems fail to extract information in a consistent and organized way. The authors alleged that there is no such approach that can be believed to be an ideal one for all the situations.

Achilonu et al. (2022) [24] put forward a rule-based NLP algorithm that extracted the unstructured data from pathology reports for breast cancer prediction and modified treatment strategy.

The previous literature reflects strong evidence about the occurrence of abnormal LFTs as well as lipid levels in a human being having diabetes mellitus.

Mandal et al. (2018) [25] applied the unpaired t-test, chi-square/fisher's exact test using the statistical package for social sciences (SPSS) to find the association between the liver enzymes of the disease dataset. The dataset showed elevation in the alanine aminotransferase levels of diabetic patients, but no such rise was seen in aspartate aminotransferase (AST), gamma-glutamyl trans peptidase (GGTP) and alkaline phosphatase (ALP) levels.

Bhowmik et al. (2018) [26] studied the rural Bangladeshi population dataset and explored the association between lipid profile with diabetes and pre-diabetes within it. The authors showed the relationship between the serum lipids and glucose intolerance status in diabetic population using the analysis of covariance (ANCOVA) and regression analysis. The dataset consisted of the 90% of low high-density lipoproteins (HDL) levels. The authors found out an essential association of pre-diabetes with high triglycerides (TG). The study observed a linear trend for high total cholesterol (TC), high TG and low HDL with increasing glucose intolerance.

Singh et al. (2019) [27]studied the diabetic dataset of north Indian population and found out the elevation in AST, ALP and bilirubin with the frequencies of 59.3%, 52.6%, 42.1% and 31.5% respectively. The frequency of diminished albumin in diabetic people was recorded as 73.6%. The study additionally indicated that the ALP concentration associated with the blood sugar fasting levels increased considerably.

Majid et al. (2019) [28] used the SPSS version 22 to analyze the LPP of the diabetic dataset and showed the abnormalities in the TG, TC, HDL and low-density lipoproteins (LDL) serums. The abnormalities included high serum TG, high serum TC, low HDL and high LDL in 58.1%, 61.9%, 44.8% and 53.3% of the dataset. The age group was found linked to the HDL levels.

Shahwan et al. (2019) [29] showed the considerable relationship between abnormal serum hepatic enzymes, glycemic control and lipid levels in diabetic patients.

Islam et al. (2020) [30] used the multinomial logistic regression analysis that showed that the serum GGTP activity increased independently of diabetes in adults. The average concentrations of AST, ALP and GGTP serum were significantly higher in the diabetic group as compared to the non-diabetes group.

Blomdahl et al. (2021) [31] found that the liver disease has the effect on the insulin resistance. The diabetic populations who consume moderate alcohol are at a higher risk of getting fibrosis.

Tham et al. (2021) [32] found out that apart from the traditional risk factors, the plasma lipids can help in the detection and prediction of the atrial fibrillation in diabetic people. The authors used the logistic regression models on the collected dataset for conducting the study.

Kosmalski et al. (2022) [33] found that there is an underestimation of the liver disease problem in diabetic people and showed the coexistence of both the diseases. The authors alleged that glucose metabolism disturbances diagnosis is typically not performed in patients having liver disease.

The PDFs do not follow a generalized structure. Moreover, it is very difficult to separate the header/footer of the PDF from the rest of its content. The challenges with the various proposed methods are mostly faced due to specific recording styles that lead to a lack of generalization. Since the unstructured clinical text in the form of PDF reports contains intricacies, consequently there is a need for scalable and high-performing approach to extract the data from the unstructured format to a structured one which can be used to enhance health outcomes. Also, there is a theoretical research gap. The research community has provided enough theory that can be applied to generate a new disease dataset where a disease can be predicted or diagnosed from the markers of another connected disease/s. The past literature ascertains the interconnection of liver, lipid and diabetes diseases. However, there is no such machine learning technique that has been implemented in terms of the predictive effectiveness of the liver enzymes and the lipid levels for diabetes detection or prediction. The literature also points to the fact that the standard datasets have been mostly used to diagnose the diabetes disease and there is a need to incorporate more significant attributes or risk

factors to determine the disease. The geographical location also affects the health of individuals; hence, new dataset can lead to new insights pertaining to a particular region.

3.Methods

With the purpose of monitoring different diseases, separate clinical lab test reports are possibly placed at different locations. Moreover, it becomes very time consuming to understand and make use of the unstructured content manually. There is a need of converting the unstructured data to structured one in order to gain the valuable insights about the data. For that purpose we have chosen PDF Scraping i.e., extracting unstructured data of LFT, LPP and glycated haemoglobin (HbA1c) from PDF documents. *Figure 1* depicts the extract, transform and load (ETL) process used in the research study as the methodology for extracting the structured data from unstructured PDFs.

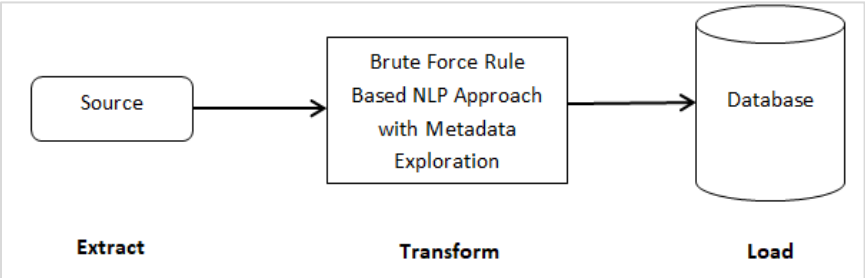


Figure 1ETL methodology

In order to arrive at a structured dataset, the unstructured data from the PDFs is transformed into the structured format using the proposed improved brute force rule-based NLP approach to metadata exploration.

3.1Data

Data collection is an organized procedure of assembling, analyzing and understanding various types of information from numerous sources [31]. The dataset used in this study is a primary clinical dataset collected from a diagnostic test centre located in Srinagar, Jammu and Kashmir. The data was collected under the expert medical supervision in the form of unstructured PDF diagnostic reports excluding the name of the patients as the privacy and anonymity of patients was taken with great care. The collected data did not follow any format when acquired. The reports consist of different disease diagnostic test reports like LFT, LPP and HbA1c test. The PDFs contain the test reports of 3350 patients of nearly all age groups. The test reports show the measures of certain enzyme and protein levels in the blood. The data was collected for a period of six months specifically from February 2021 to August 2021. *Table 1* shows the description of the population used for the study.

Table 1 Population description

Population percentage		**Age group**		**Children**		**Adolescents**		**Adults**		**Old Adults**	
Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
50.76	49.24	9-72	13-79	22	2	545	831	867	578	262	234

3.2Data structure

The small set of basic data object types form the building blocks of a PDF document - a data structure. Three PDF file data structures are used to perform the exploration of the diabetes disease dataset and their details are along the lines below:

3.2.1Liver function test PDF lab report

Figure 2 shows the LFT lab report, which consists of eleven attributes. These are bilirubin total, direct bilirubin, indirect bilirubin, total protein, albumin, globulin, a/g ratio, sgot, sgpt, alkaline phosphatase and gamma gt. The report also contains the demographic attributes such as age and gender of the patient. The normal values for the liver enzymes are specified in the report against each test.

Eight out of eleven attribute values are acquired by the direct measurement using the specified tests mentioned in the report. The indirect bilirubin, globulin and a/g ratio are calculated from the already measured constituent tests of LFT. The indirect bilirubin value is given by the difference between the

values of bilirubin total and direct bilirubin. The globulin value is calculated as the difference between values of total protein and albumin while the ratio of albumin and globulin values give a/g ratio.

Age/Gender: 50 Y/Male
Patient ID: 012011140023
BacodeNo: 10199272

BIOCHEMISTRY
LIVER FUNCTION TEST(LFT)

Test Name	Value	Unit	Bio Ref.Interval
Specimen: Serum			
BILIRUBIN TOTAL DPD	1.2	mg/dL	0.1-1.2
DIRECT BILIRUBIN(CONJUGATED), Serum DPD	0.2	mg/dL	< 0.20
INDIRECT BILIRUBIN(Unconj.),Serum Calculated	**1.0**	mg/dL	0.80
TOTAL PROTEIN , Serum Biuret	8.0	g/dL	6.6-8.3
ALBUMIN,SERUM BCG	4.80	g/dL	3.5-5.0
GLOBULIN,Serum Calculated	3.2	g/dL	2.3-3.5
A/G Ratio ,Serum Calculated	1.50		1.0 - 2.3
SGOT (AST) ,Serum IFCC	30.00	U/L	< 50.0
SGPT (ALT), Serum IFCC	48.00	U/L	< 50
ALKALINE PHOSPHATASE ,Serum IFCC	**145.0**	U/L	30-120
GAMMA GT ,Serum IFCC	46.00	U/L	5 - 64

Figure 2 Snapshot of LFT lab report

3.2.2Lipid profile panel PDF lab report

Figure 3 shows the LPP lab report that contains seven attributes. These are total cholesterol, triglyceride, hdl cholesterol, ldl cholesterol, vldl, total cholesterol/hdl ratio and ldl/hdl cholesterol ratio. Demographic attributes such as age and gender of the patient are mentioned in the report. The normal values for the lipid panel are specified in the report. Three out of seven attribute values are acquired by the direct measurement using the specified tests mentioned in the report. The rest of the four tests constituting LPP are calculated. The ldl cholesterol is given by dividing the triglyceride value by five and then subtracting it from the difference of total cholesterol and hdl cholesterol values. The vldl is given by dividing the triglyceride value by five. The total cholesterol/hdl ratio is given by dividing the total cholesterol value with the hdl cholesterol value, whereas ldl/hdl cholesterol ratio is given by dividing the ldl cholesterol value with hdl cholesterol value.

Age/Gender: 50 Y/Male
Patient ID: 012011140023
BacodeNo: 10199272

BIOCHEMISTRY
LIPID PROFILE

Test Name	Value	Unit	Bio Ref.Interval
TOTAL CHOLESTEROL Enzymatic(CHO-POD)	191.00	mg/dL	Desirable < 200 Borderline 200 - 239 High > 240
TRIGLYCERIDE , Serum GPO-POD	**288.00**	mg/dL	Normal 150 Border line high 150-199 High 200-490 very High> 500
HDL-CHOLESTEROL , Serum Direct measure	**38.00**	mg/dL	40-60
LDL CHOLESTEROL,Serum Calculated	95.40	mg/dL	Optimal<100 Near or Above Optima-l00-129 Borderline High 130 - 159 High 160 - 189 Very High >190
VLDL ,Serum Calculated	**58.00**	mg/dL	0.0- 30.0
TOTAL CHOLESTEROL /HDL RATIO ,Serum Calculated	**5.03**		Indicates low Risk < 3.0 Indicates Average Risk3.0-5.0 Indicates High Risk > 5.0
LDL / HDL CHOLESTEROL RATIO Calculated	2.51		1.5-3.5

Figure 3 Snapshot of LPP lab report

3.2.3Glycated haemoglobin profile PDF Lab Report

Figure 4 shows the HbA1c lab report that consists of two attributes. The attributes are glycosylated HbA1c and estimated average glucose. The normal range of values is specified in the report against each test. The demographic attributes such as age and gender of the patient are mentioned in the report. Both test attribute values are acquired by the direct measurement. The result of this test report will be taken as the decisive factor in diabetes disease diagnosis and prognosis and hence will serve as a target variable of the dataset.

Age/Gender: 50 Y/Male
Patient ID: 012011140023
BacodeNo: 10199272

HAEMATOLOGY
HbA1C

Test Name	Value	Unit	Bio Ref.Interval
Specimen: Whole Blood EDTA			
Hb A1C, GLYCOSYLATED Hb by HPLC	10.00	%	Non-Diabetic < 6.0 Good Control 6.0-7.0 Weak Control 7.0-8.0 Poor control > 8.0
Estimated Average Glucose	240.30	mg/dL	68-125

Figure 4 Snapshot of glycated hemoglobin lab report

3.3Algorithm

PDF scraping is moderately comparable to text analytics and with this method worthy information can be derived from the text. Similarly, we tried to extract the clinical or diagnostic values of test reports from the PDF reports of patients using a rule-based approach and the process automatically extracts the disease test values from the reports. *Figure 5* shows the flow of our proposed approach to convert unstructured data to structured one.

The PDF files are stored in a folder and then the reports are parsed one by one automatically. Since page headers and footers are not separated from the rest of the page content, it becomes very difficult to extract them from PDFs. We proposed to use the metadata of the PDF to estimate the headers and/or footers part. The dictionary contents are retrieved precise information about the font properties used in the PDF. Using the font information, we have separated the headers, footers and other auxiliary information contained in the PDFs. Our scraping technique for each PDF document involves the use of a Python library named "PyMuPDF". The library also supports the standard metadata of the PDFs. *Figure 6* shows the pseudo code for extracting each page of the PDF document. Once the whole PDF gets parsed, the brute force rule based algorithm is applied on the resultant intermediate text file to extract the required column name and its value. *Figure 7* shows the pseudo code for the pattern matching algorithm used in this research. The column name and the value are split into the key-value pair and then converted to the dictionary. With the help of "pandas", the dictionary is converted to the data frame which is then exported to comma-separated values (CSV).

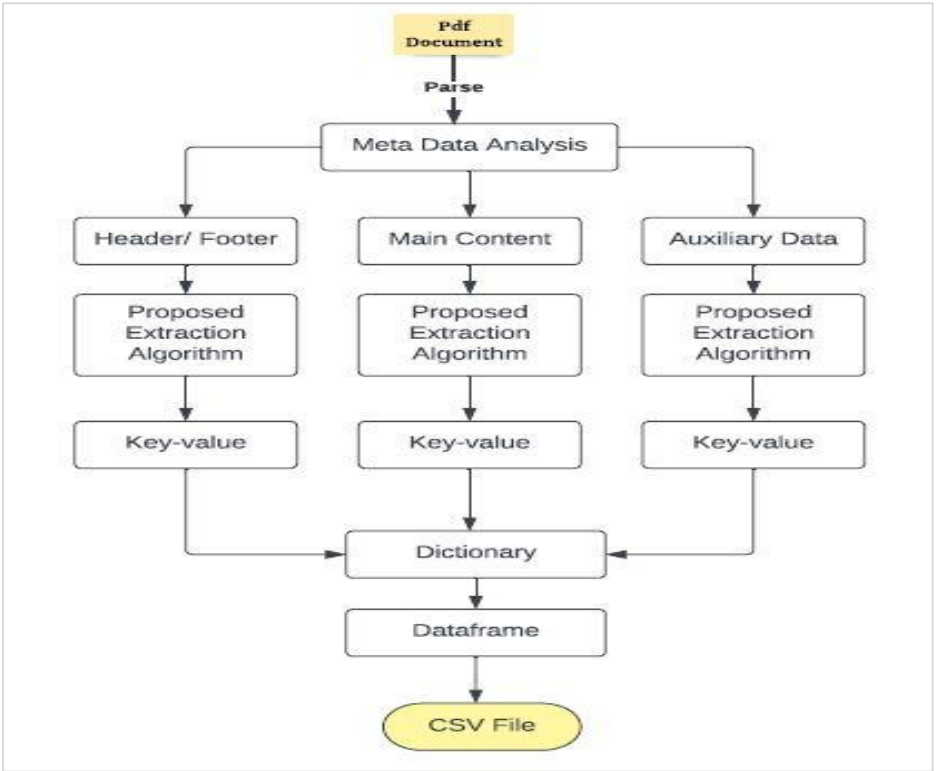


Figure 5 Flow diagram of proposed approach

```
Extract_Page_No (Total_Length -1)
{
   FOR each Page_No in range 0 to Total_Length -1
   Page ← Doc (Page_No)
   Words ← Page.GET_TEXT (Words)
   FOR A in range Page.ANNOTS ( )
   IF A!= NONE
   Rectt = ANNOT.Rect
   pages ← w for w in WORDS if fitz.Rect (w [:4]) in Rect
   ANN ← MAKE_TEXT (pages)
   ALL_ANNOTS.append (ANN)
}
```

Figure 6 Pseudocode for PDF parsing

```
Pattern_Matching (VAR_1  & VAR_2)
{
  L_VAR_1 ← Len (VAR_1)
  L_VAR_2 ← Len (VAR_2)
  MAX ← (L_VAR_1 - L_VAR_2 +1)
  FOR Pat_Mat_1 in Range 1 to MAX
    SET Flag to TRUE
    FOR Pat_Mat_2 in Range 1 to L_VAR_2
      IF VAR_2 [Pat_Mat_2]!= VAR_1  [Pat_Mat_2 + Pat_Mat_1 -1]
       SET Flag to FALSE
      IF Flag == TRUE
       Return Pat_Mat_1
  ELSE Return 0
}
```

Figure 7 Pseudocode for Pattern Matching

3.4Implementation

The proposed approach has been implemented in the Python programming language using the PyMuPDF and Pandas libraries. The data analytics tool used for the implementation is the Jupyter Notebook.

3.5Performance analysis

Many performance parameters are used to evaluate the text extraction technique. Some of them are given in Equation 1, Equation 2 and Equation 3:

$$Precision\ Rate = \frac{\text{correctly detected}}{\text{correctly detected + false positive}} \quad (1)$$

$$Detection\ Rate = \frac{\text{correctly detected}}{truth\ text} \quad (2)$$

$$Recall\ Rate = \frac{\text{correctly detected}}{\text{correctly detected+false negative}} \quad (3)$$

4.Results

After the application of the proposed algorithm, the database is formed containing the predictive attributes from the liver and lipid profile of the patients. The target variable for each patient record is set using their HbA1c test value. The target class is categorized into three values – normal, prediabetic and diabetic. The final database formed is clean and can be directly used for the data analysis. The collected diabetes database contained 3350 records. Out of the 3350 records, 33 records had missing or misappropriate data. *Figure 8* shows the snapshot of the database formed. The proposed approach that we have used to extract, integrate and formulate a structured data format uses least computing resources. Querying 3350 clinical PDF reports took, at the most 5 minutes. Precisely 11 documents were parsed and extracted in 1 second i.e., 670 documents were parsed in a minute. The usable field rate was calculated to be 3298/3350 = 98.44%. This rate is quite remarkable as compared to the rule based approaches used in the past. A total of 23 rules was formed to extract the unstructured data and convert it into a structured format.

age	gender	t_bilirbn	d_bilirbr	i_bilirbn	tpro	alb	glb	agr	sgot	sgpt	alp	ggtp	tchol	tglyc	hdl	ldl	vldl	tc_hdl	ldl_hdl	target
47	Female	0.8	0.1	0.7	7.9	4.5	3.4	1.32	17	16	86	35	130	197	42	48.6	39	3.1	1.16	NORMAL
40	Male	1.5	0.3	1.2	8.3	4.8	3.5	1.37	33	53	145	27	106	121	40	41.8	24	2.65	1.05	NORMAL
35	Female	0.4	0.1	0.3	8	4.6	3.4	1.35	26	12	62	10	156	276	40	60.8	55	3.9	1.52	NORMAL
42	Female	1.3	0.2	1.1	7.6	4.7	2.9	1.62	15	10	65	12	184	83	50	117.4	17	3.68	2.35	NORMAL
70	Female	0.4	0.1	0.3	7.9	4.5	3.4	1.32	35	29	88	77	183	254	60	72.2	51	3.05	1.2	DIABETIC
70	Female	0.9	0.2	0.7	8	4.6	3.4	1.35	25	26	105	12	202	159	52	118.2	32	3.88	2.27	PREDIABETIC
68	Male	0.7	0.2	0.5	8.3	4.9	3.4	1.44	17	17	117	27	156	107	45	89.6	21	3.47	1.99	DIABETIC
30	Female	0.9	0.2	0.7	8	4.9	3.1	1.58	35	47	86	10	168	73	58	95.4	15	2.9	1.64	NORMAL
59	Female	0.5	0.1	0.4	8	4.5	3.5	1.29	19	19	95	22	198	128	49	123.4	26	4.04	2.52	DIABETIC
35	Male	0.9	0.2	0.7	8	5	3	1.67	28	20	62	31	172	100	62	90	20	2.77	1.45	NORMAL
21	Male	0.5	0.1	0.4	8.2	5	3.2	1.56	36	41	108	67	146	176	42	68.8	35	3.48	1.64	NORMAL
41	Male	0.8	0.1	0.7	7.4	4.4	3	1.47	28	26	86	24	189	167	41	114.6	33	4.61	2.8	NORMAL
52	Male	1.9	0.5	1.4	7.4	4	3.4	1.18	28	30	106	10	120	51	41	68.8	10	2.93	1.68	PREDIABETIC
30	Female	0.5	0.1	0.4	8	4.7	3.3	1.42	22	17	103	11	186	98	63	103.4	20	2.95	1.64	NORMAL
38	Male	2.2	0.3	1.9	8.2	4.7	3.5	1.34	62	89	117	59	213	260	47	114	52	4.53	2.43	PREDIABETIC
21	Female	0.5	0.1	0.4	7.6	4.3	3.3	1.3	37	34	82	10	154	110	60	72	22	2.57	1.2	PREDIABETIC
32	Male	0.6	0.1	0.5	7.7	4.7	3	1.57	94	179	91	101	202	148	46	126.4	30	4.39	2.75	PREDIABETIC
55	Female	1	0.2	0.8	7.8	4.6	3.2	1.44	25	22	108	20	183	149	51	102.2	30	3.59	2	NORMAL
56	Male	0.9	0.2	0.7	7.6	4.6	3	1.53	19	25	81	84	157	97	47	90.6	19	3.34	1.93	PREDIABETIC
32	Female	0.8	0.1	0.7	7.4	4.4	3	1.47	37	81	78	33	149	137	49	72.6	27	3.04	1.48	NORMAL
75	Female	0.4	0.1	0.3	7.5	4	3.5	1.14	39	33	155	14	147	113	50	74.4	23	2.94	1.49	DIABETIC
50	Male	0.6	0.1	0.5	7.8	4.6	3.2	1.44	35	53	98	152	179	158	48	99.4	32	3.73	2.07	NORMAL
46	Male	0.7	0.1	0.6	8.2	4.7	3.5	1.34	37	57	77	52	229	524	49	75.2	105	4.67	1.53	DIABETIC
46	Male	2	0.3	1.7	7.8	5	2.8	1.79	205	213	74	55	144	190	40	66	38	3.6	1.65	NORMAL
45	Male	1.6	0.2	1.4	7.4	4.5	2.9	1.55	32	52	124	85	136	104	46	69.2	21	2.96	1.5	DIABETIC
33	Male	0.6	0.1	0.5	7.8	4.6	3.2	1.44	48	87	85	102	183	195	40	104	39	4.58	2.6	PREDIABETIC
21	Male	1.6	0.2	1.4	8.1	5	3.1	1.61	51	108	106	26	173	191	40	94.8	38	4.33	2.37	NORMAL
38	Female	0.7	0.1	0.6	8.1	4.8	3.3	1.45	21	19	94	15	162	96	57	85.8	19	2.84	1.51	NORMAL

Figure 8 Snapshot of populated diabetes database

Table 2 gives the calculated values of various performance parameters used in the study to evaluate the proposed brute force rule based NLP technique.

Table 2 Performance of proposed approach

Parameters	Value
Precision Rate	98.94%
Detection Rate	98.44%
Recall Rate	99.48%

4.1Value of the data

The collected diabetes dataset contains 3350 records with 1645 female patient records and 1696 male patient records (*Figure 9*). The dataset contains 20 predictive attributes and one target attribute. The predictive attributes contain 11 attributes or features from the LFT and 7 from the LPP. The two features are demographic in nature. The target class was labeled using the outcomes of HbA1c test. The target class is divided into three classes on the basis of the results of HbA1c test. The three classes are normal, prediabetic and diabetic. The target class which is categorized as well as discrete in nature has 1783 normal patient records, 795 prediabetic patient records and 772 diabetic patient records. The visual description of the target class of diabetes dataset in terms of disease distribution is shown in *Figure 10*.

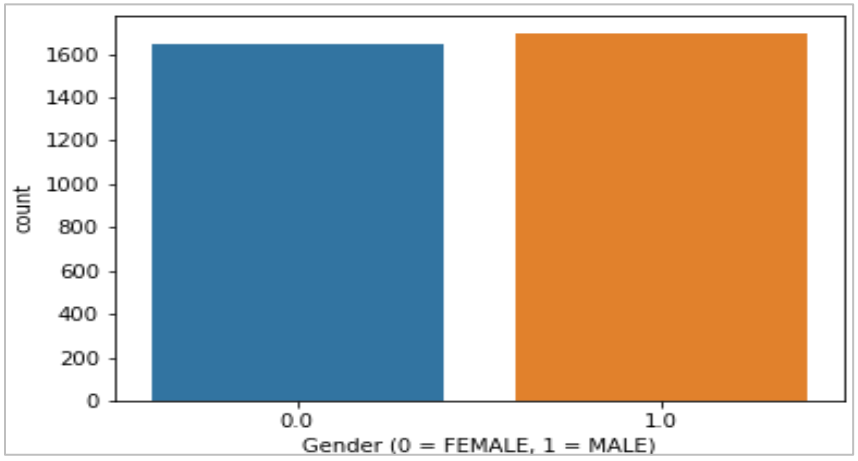


Figure 9 Barplot showing gender distribution in extracted dataset

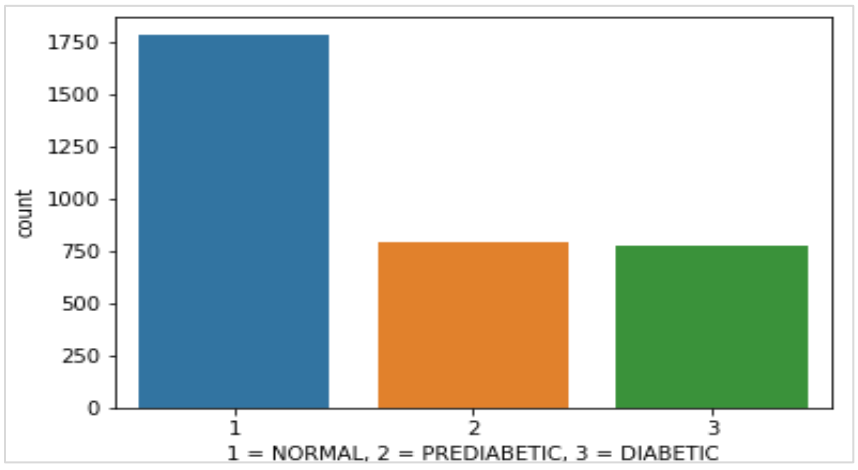


Figure 10 Barplot showing target class distribution in extracted dataset

The data reflect the liver profile and LPP diagnostic tests in Srinagar, Jammu and Kashmir. It is possible to analyze the effects of LFT and LPP diagnostic tests on the diabetic people in Srinagar, Jammu and Kashmir, based on it, it is possible to draw conclusions about the prediction accuracy of the model. The acquired data can be used to predict the diabetes disease from the hepatic [34] as well as lipid panel and also the major risk factors can be identified that affect the diabetes disease. The data will enhance

opportunities for the diabetes prediction and can be used for comparing the predictions with those given by standard benchmark diabetes datasets.

4.2Data exploration

The disease data records are integrated from different PDF test reports of the same patient, organized on the basis of the patient id and subsequently the database is populated. *Table 3* shows the feature set description of the collected and extracted diabetes dataset. It evidently shows that the majority of the values have a right skewed distribution. The skew is present because of the disproportional dispersion of the data. The features of the dataset are either directly or indirectly dependent on each other. They share a correlation among themselves that acts as a common tool for describing their relationships. Correlation is a statistical measure [35] that states the range to which two variables are linearly associated. The correlated features change together at a constant rate. *Table 4*, *Table 5* and *Table 6* show the ascending order of the correlation of the attributes of the diabetes dataset calculated by using Pearson correlation coefficient [36], Spearman's rank correlation coefficient [37] and Kendall's rank correlation [38] respectively.

Table 3 Feature set distribution of the extracted dataset

Attribute	Description	Distribution	Mean	Median	Mode	Standard deviation	Variance
		Numerical values					
age	Age in years	normal	44.02	42.00	45.00	14.15	200.22
t_bilirbn	Total Bilirubin	right skewed	0.88	0.80	0.60	0.49	0.24
d_bilirbn	Direct Bilirubin	right skewed	0.17	0.20	0.10	0.11	0.01
i_bilirbn	Indirect Bilirubin	right skewed	0.70	0.60	0.50	0.40	0.16
tpro	Total Protein	left skewed	7.82	8.00	8.00	0.38	0.14
alb	Albumin	left skewed	4.59	4.60	4.50	0.34	0.11
glb	Globulin	left skewed	3.23	3.30	3.30	0.22	0.04
agr	Albumin/Globulin Ratio	normal	1.42	1.41	1.29	0.16	0.02
sgot	Serum Glutamic - Oxaloacetic Transaminase	right skewed	35.72	30.00	23.00	26.94	725.76
sgpt	Serum Glutamic Pyruvic Transaminase	right skewed	43.05	33.00	18.0	34.63	1199.23
alp	Alkaline Phosphatase	right skewed	111.70	105.00	105.00	39.86	1588.82
ggtp	Gamma-Glutamyl Transferase	right skewed	37.07	25.00	14.00	39.66	1572.91
tchol	Total Cholesterol	normal	181.03	180.00	189.00	40.29	1623.28
tglyc	Serum Triglyceride	right skewed	170.69	148.00	121.00	85.86	7371.94
hdl	High-Density Lipoprotein	normal	46.90	45.00	41.00	8.37	70.05
ldl	Low-Density Lipoprotein	normal	99.97	97.80	69.40	32.31	1043.93
vldl	Very-Low-Density Lipoprotein	right skewed	34.04	30.00	24.00	17.21	296.18
tc_hdl	Total Cholesterol/High-Density Lipoprotein Ratio	normal	3.90	3.85	3.00	0.79	0.62
ldl_hdl	Low-Density Lipoprotein/High-Density Lipoprotein Ratio	normal	2.15	2.09	2.04	0.65	0.42
		Categorical values					
gender	Male or Female	binary	0.50	1.00	1.00	0.50	0.25
target	Target Variable - Test for Diabetes	binary	1.69	1.00	1.00	0.81	0.65

Table 4 Pearson correlation with target class in ascending order

Attribute	Pearson correlation with target class
target	1.000000
age	0.352351
ggtp	0.176638
tglyc	0.140428
vldl	0.139001
agr	0.080002
sgot	0.078362

Attribute	Pearson correlation with target class
alp	0.075641
glb	0.072712
i_bilirbn	0.053219
sgpt	0.047023
alb	0.045465
t_bilirbn	0.044464
ldl	0.040396
tc_hdl	0.038506
ldl_hdl	0.036338
hdl	0.033244
tchol	0.020587
gender	0.016169
d_bilirbn	0.005724
tpro	0.001787

Table 5 Spearman correlation with target class in ascending order

Attribute	Spearman correlation with target class
target	1.000000
age	0.364305
ggtp	0.249409
tglyc	0.149587
vldl	0.146996
alp	0.099404
agr	0.066746
glb	0.063636
sgot	0.062626
tc_hdl	0.055735
i_bilirbn	0.052474
sgpt	0.050649
t_bilirbn	0.042001
alb	0.040269
tchol	0.036291
hdl	0.031999
ldl	0.030949
ldl_hdl	0.025460
tpro	0.016946
d_bilirbn	0.010427
gender	0.008719

Table 6 Kendall correlation with target class in ascending order

Attribute	Kendall correlation with target class
Target	1.000000
Age	0.289290
Ggtp	0.195804
Tglyc	0.116311
Vldl	0.115238
Alp	0.077781
Agr	0.052981
Glb	0.052966
Sgot	0.049902
tc_hdl	0.043390
i_bilirbn	0.043107
Sgpt	0.039679
t_bilirbn	0.034158
Alb	0.032936
Tchol	0.028491

Attribute	Kendall correlation with target class
Hdl	0.025538
Ldl	0.023735
ldl_hdl	0.019737
Tpro	0.013961
d_bilirbn	0.009304
Gender	0.008290

4.3Data visualization

Data visualization is the demonstration of data or information [39] in a visual format. It has become very imperative from the analytics viewpoint. Visualization is a tool for explainable machine learning for exploratory data analysis. It acts as a knowledge generator helping the analyst to draw hypothesis about the observed data from model outcomes [40]. The visual elements graphically represent the information and data in order to see and understand the outliers and patterns it contains.

The graphical representation of data can range from single to multiple dimensions. Many data visualization tools as well as visual elements like graphs, charts, and maps help to understand the trends and patterns followed by the data. The most commonly used visualization methods and plots in machine learning are line plot, scatter plot [41], box-plot [42], etc. *Figure 11* shows the bar plot of the first 200 records of the dataset. The bar plot used visualizes the comparisons between the three target classes i.e., normal, prediabetic and diabetic on the basis of the patient's age.

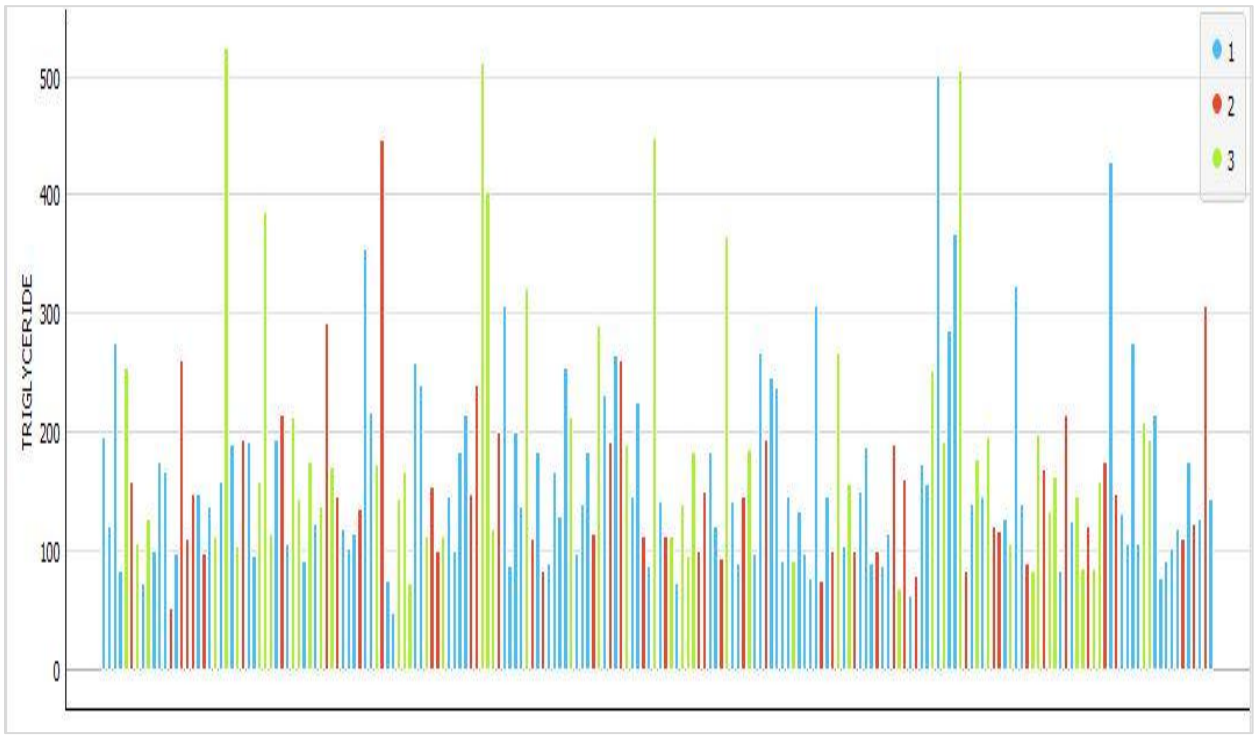


Figure 11 Bar plot comparison of three target classes on the basis of TG (1-Diabetic, 2- Normal, 3 - Prediabetic)

Figure 12 also shows the bar plot comparison of male and female patients on the basis of TG. *Figure 13* shows the scatter plot of LDL cholesterol versus TC and patient age versus GGTP respectively. These plots assist in viewing and observing relationships between two numeric variables [41].

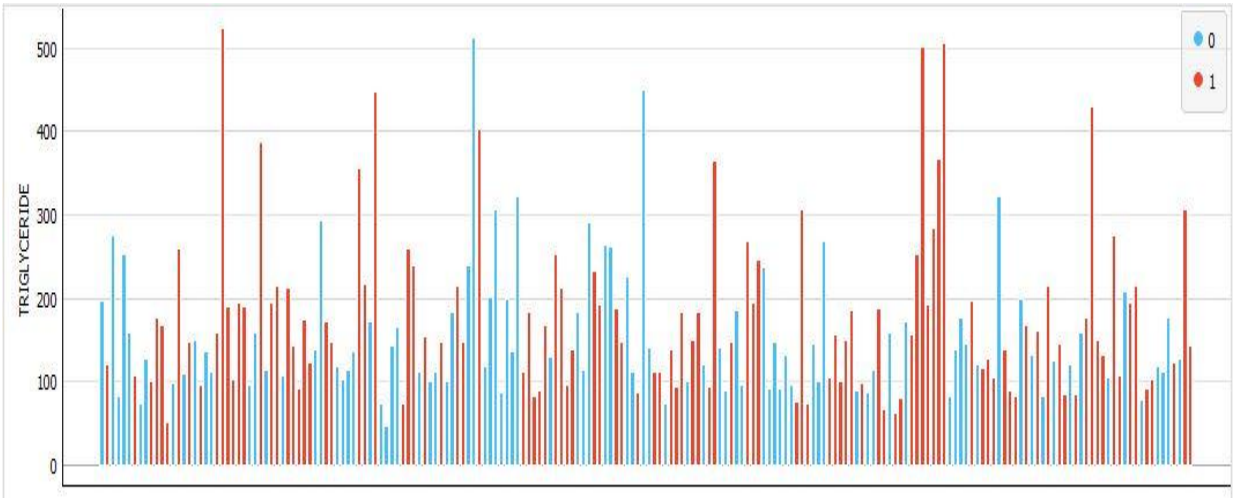
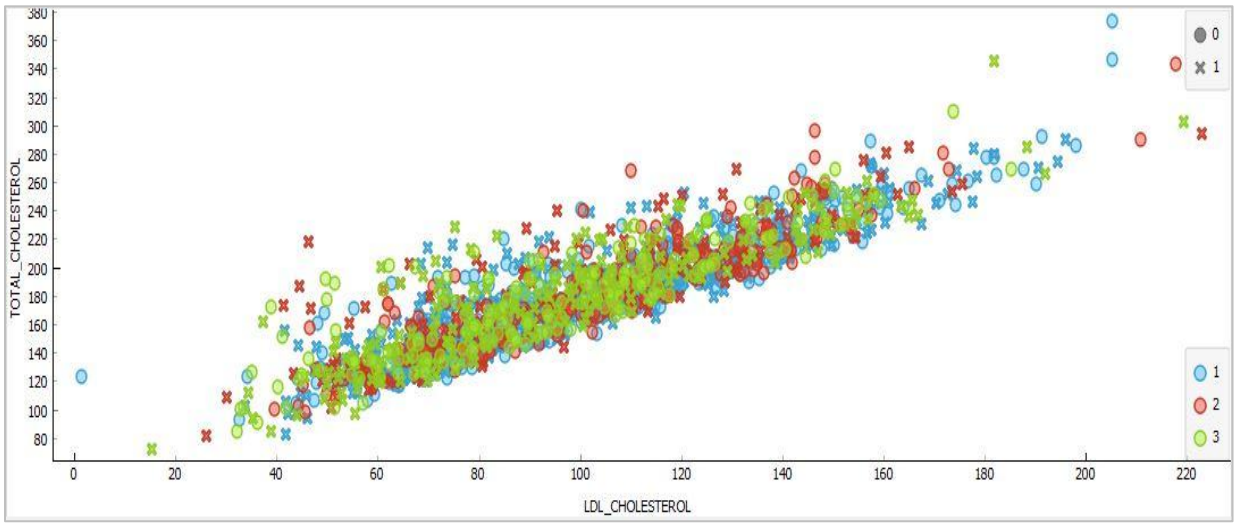
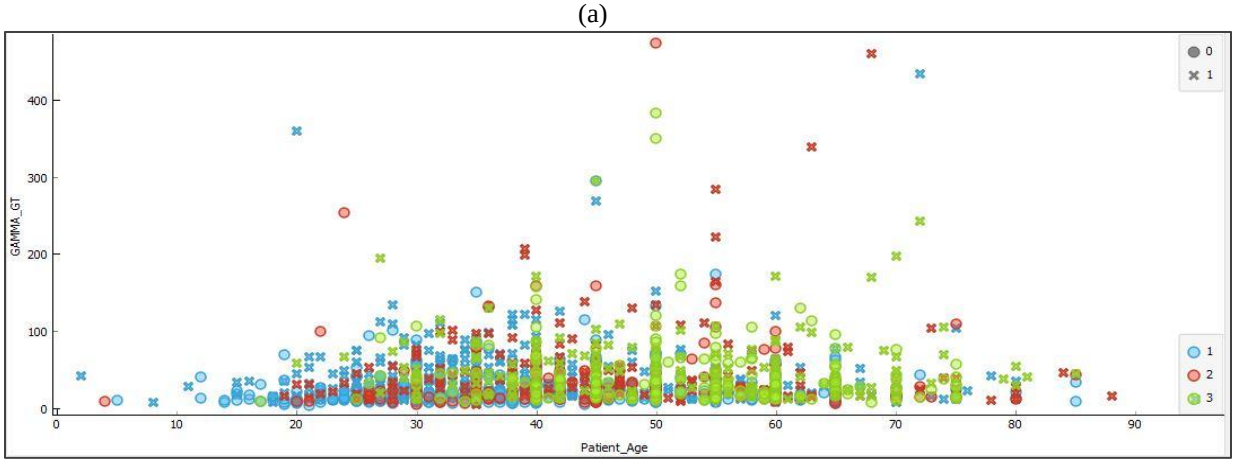


Figure 12 Bar plot showing comparison between TG and gender (0- Female, 1-Male)



(a)



(b)

Figure 13 Scatter plot of (a) LDL cholesterol versus TC (b) patient age versus GGTP (1-Diabetic, 2- Normal, 3 – Prediabetic; 0-Female, 1- Male)

The collected dataset contains outliers; therefore, it becomes very difficult to visualize the data. Also, the presences of outliers and un-scaled data tend to slow down or degrade the predictive performance of the algorithms. It becomes very important to switch to feature scaling which considers on converting the values into the similar range or similar scale. In order to provide non-linear transformations, quantile transformers or scalers [43] are used. These scalers try to shrink the distances between the marginal outliers and inliers. In machine learning, feature scaling is regarded as one of the most critical steps during the pre-processing [44] of the data. The performance of a machine learning model highly depends upon this step. The quantile scaling tries to make the two distributions identical in statistical properties and is used to suppress the intervention of outliers.

Figure 14 shows the quantile scaled heatmap [45] of the dataset. Heatmap is a data visualization tool where the phenomenal magnitude is represented by colours. It is a two dimensional representation of data and is used in data analytics. For comparing categorical with continuous data, there are other visualization methods available. One such method is Mosaic display. It visualizes the categorical data over a pair of variables [46]. *Figure 15* shows the Mosaic display of gender versus age in normal, pre-diabetic and diabetic patients of the dataset. For considering the pattern of association among categorical variables, another type of visualization method known as Sieve diagram [47] is used. *Figure 16* shows the Sieve diagram of the age versus the target values, where both the predictive variables are categorical in nature.

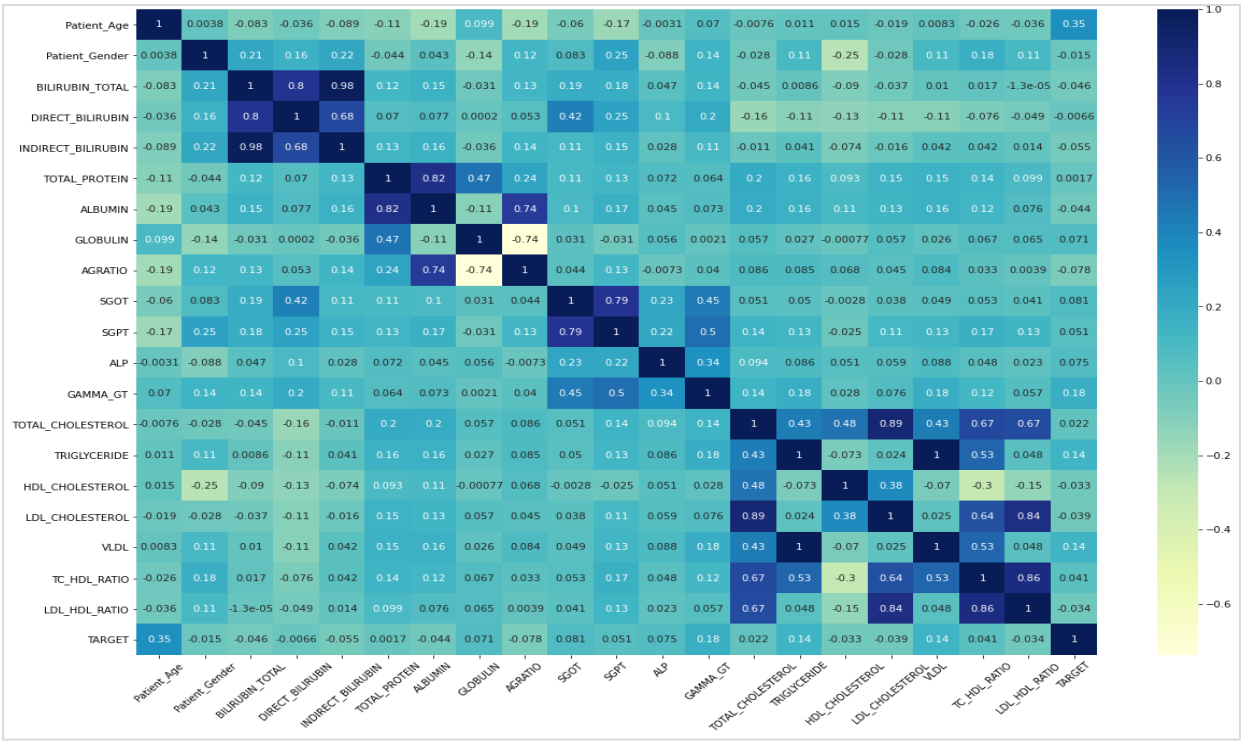


Figure 14 Heatmap for extracted diabetes dataset with quantile scaling

The graphical representation of data can range from single to multiple dimensions. It helps to understand the trends followed by the data. For example, in order to create the graphical content, 3D visualization is used. It familiarizes an object in the form of an image in three-dimensional space. 3D visualization (*Figure 17*) is performed on the attributes sgpt from LFT report, tchol from LPP report and age (demographic attribute) contained in the created database. The outliers present in the 3D scatter plot are because the liver enzymes and lipid levels escalate or decrease from their normal range as a result of liver damage and lipid disorder in diabetic patients. *Figure18* shows the multivariate visualization approach of the dataset using the FreeViz [48] projection of the data. Using the same graph, these visualizations present data on many features. The projection that best separate instances of different class are chosen through an optimization procedure [49].

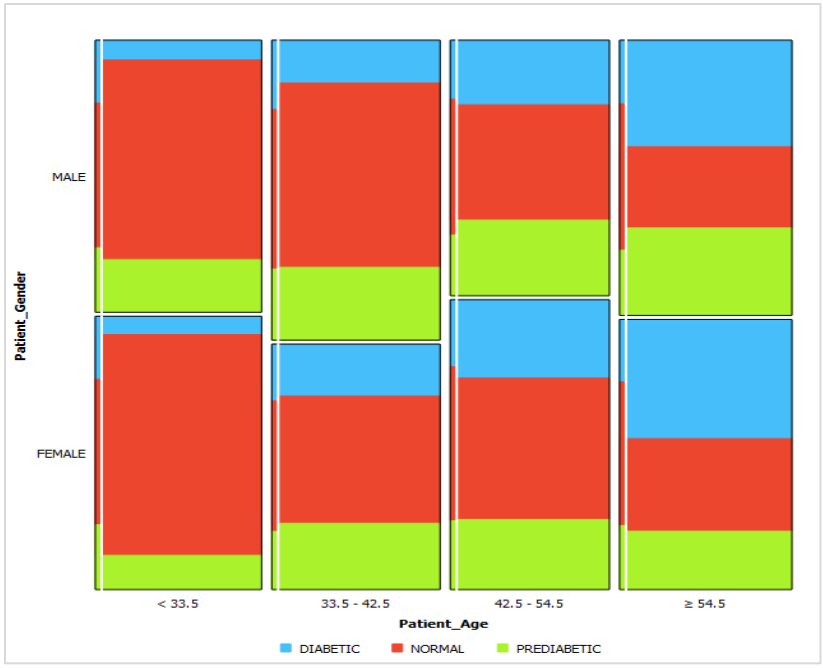


Figure 15 Mosaic display of gender versus age in normal, pre-diabetic and diabetic patients

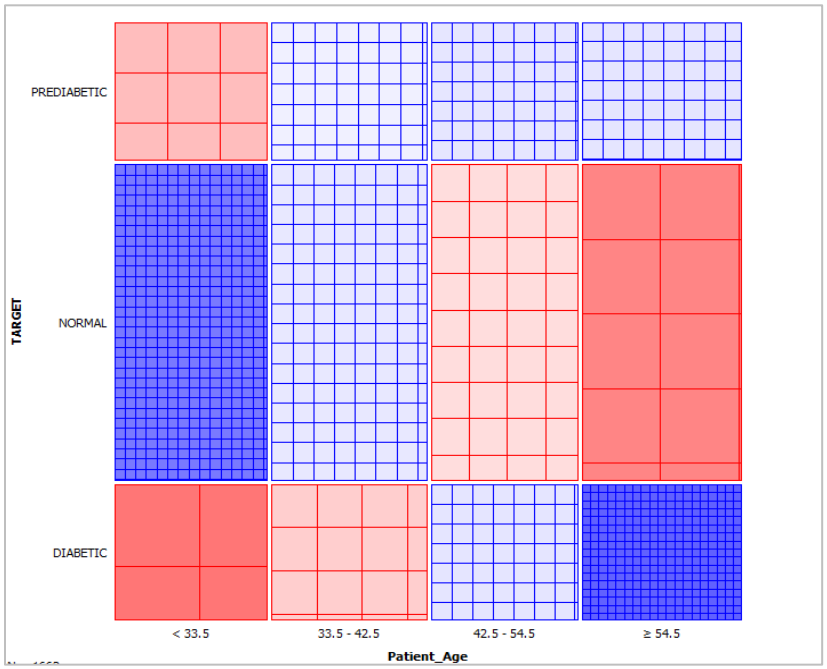


Figure 16 Sieve diagram of the age versus the target values

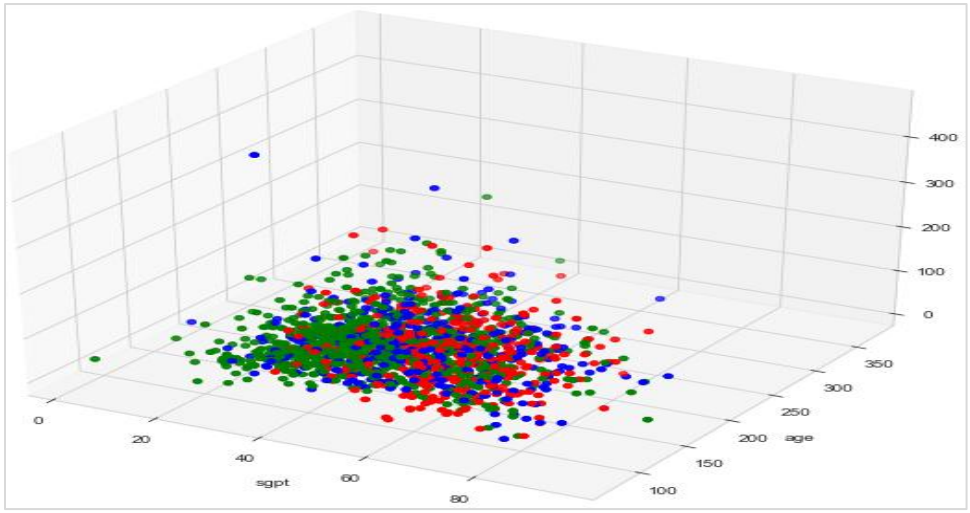


Figure 17 3D feature space of the unscaled extracted diabetes dataset

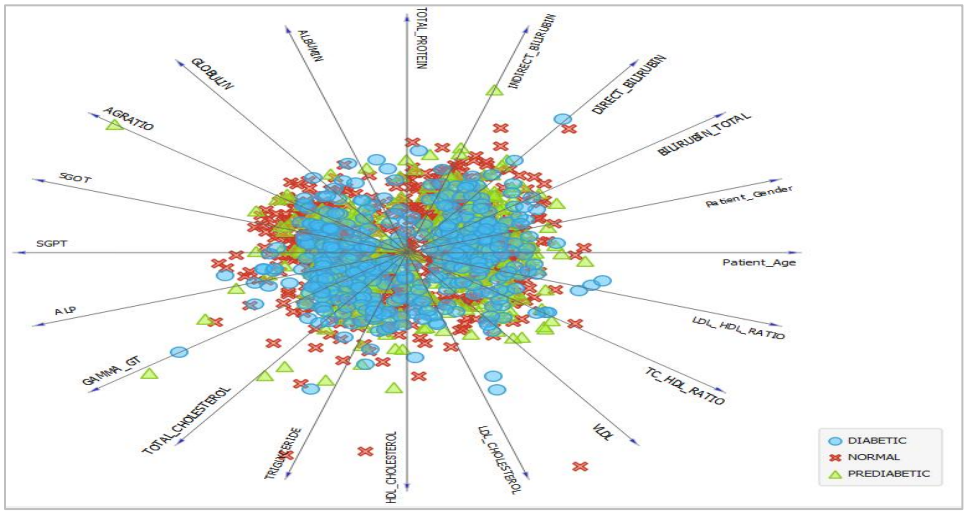


Figure 18 FreeViz projection of data

5.Discussion

Since the extraction of data from documents that have a complex structure or do not follow a particular standard structure is very time consuming. We presented an improved method in the study that extracts the data automatically using minimal computing resources. The structure in the PDF never remains same; therefore it is very difficult to find an ideal rule based approach that fits all the situations. Various extracting techniques have been presented in the literature before, but all the techniques do not fully consider the internal structure of the PDF file. The files are simply parsed without having the knowledge about the different sections of the PDF which can be worked upon. The metadata of the PDFs contain essential information regarding the document. One such information is regarding the fonts. This metadata was exploited and used to categorize the different sections of the document. Then the pattern matching technique was applied separately to the sections and the useful data was extracted. Our proposed method increased the accuracy of the usable field rates in the extracted structured format. When used without the metadata analysis, the brute force rule based NLP method performs a bit stumpy. *Table 7* shows that our proposed method with meta-analysis achieved a remarkable detection rate as compared to the other techniques. Even if a pattern which can be found in more than one section can be dealt properly and hence our technique avoids the ambiguity of the presence of the pattern in more than one section. Metadata helps in modeling the structure of the PDFs.

Table 7 Comparison of Various Extraction Techniques

Technique	Usable field rate
Regex rule based NLP approach	97.56%
Brute force rule based NLP approach	98.2%
Brute force rule based NLP approach with metadata exploration	98.44%

The formulated database can help the researchers categorize and recognize the risk factors related with the diabetes disease that can speed up its diagnosis or prognosis. The results show the correlation analysis of the risk factors that can be very helpful in determining the most important factors that affect the diabetes disease if one is suffering from liver or lipid abnormality or even both. The dataset can be used to determine the change in one attribute or enzyme with respect to the other. The dataset is helpful from both medical as well as non-medical point of view. The pragmatic implications of this research study suggest that interesting and useful patterns from the dataset can be obtained in order to predict the diabetes disease with respect to the geographical area.

The visualizations shown have the main purpose of providing the valuable insights. These visuals shown in this manuscript highlight the trends as well as the outliers that the collected dataset contains. These visualizations can allow us to gain understanding about our data by recognizing the new patterns and errors in the data. The novel idea of using LFT and LPP tests in order to predict and detect the diabetes disease can be used to translate the datasets and metrics into charts, graphs and other visuals so as to easily identify and share real-time trends and outliers. The formed dataset can definitely provide new insights about the information represented in the data with the help of visualizations. Further, this dataset can be of utmost importance to check the performance of the various machine learning techniques.

5.1Limitations

Our approach of the extraction of structured data from the unstructured documents is purely based upon the rules as well as the metadata of the document. A detailed analysis of the document is required prior to the formulation of unique string identifier tags on which the heuristics are formed upon. The rules are to be mentioned manually for querying the back end relational database. To include more clinical test reports, more rules are to be formulated in accordance with the particular test in order to populate the database.

A complete list of abbreviations is shown in *Appendix I.*

6.Conclusion and future work

This research paper presents approaches to extract, integrate and form a database of disease indicators for information discovery that helps to bridge the theoretical gap in the prediction of diabetes disease. The data is extracted from multiple unstructured PDF files; the test reports of different patients are integrated together on the basis of the patient identity number and then the database is finally populated with the diagnostic test values of liver function, lipid profile and HbA1c test using the metadata as well as heuristics formed on the basis of the unique string identifier tags found in the clinical PDF reports. The visualizations presented in the paper give a rough idea about the meaningful knowledge that can be extracted from the unstructured patient records. The knowledge can be put to use to design various diagnostic as well as predictive models. The data can be used to study the impact of diabetes on the liver enzymes as well as lipid levels of a patient and wonderful results can be inferred. The final database is clean, properly arranged, assimilated and eligible to be used for further analysis. In future we will explore the data using many descriptive as well as exploratory data analysis techniques to get further beneficial insights from the populated dataset.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

Author's contribution statement

Ifra Altaf: Conceptualization, methodology, writing - original draft and data collection. **Muheet Ahmed Butt:** Supervision, framework of methodology and interpretation of results. **Majid Zaman:** Supervision, critical feedback to shape the manuscript.

References

[1] Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications. 2011; 17(8):43-8.

[2] Natarajan Y, Kannan S, Mohanty SN. Survey of various statistical numerical and machine learning ontological models on infectious disease ontology. Data Analytics in Bioinformatics: a Machine Learning Perspective. 2021: 431-42.

[3] Taylor-weiner A, Pokkalla H, Han L, Jia C, Huss R, Chung C, et al. A machine learning approach enables quantitative measurement of liver histology and

disease monitoring in NASH. Hepatology. 2021; 74(1):133-47.
[4] Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in the diagnosis of COVID-19: challenges and perspectives. International Journal of Biological Sciences. 2021; 17(6).
[5] Rehman A, Iqbal MA, Xing H, Ahmed I. COVID-19 detection empowered with machine learning and deep learning techniques: a systematic review. Applied Sciences. 2021; 11(8):1-21.
[6] Bhavsar KA, Abugabah A, Singla J, AlZubi AA, Bashir AK. A comprehensive review on medical diagnosis using machine learning. Computers, Materials and Continua. 2021; 67(2):1997-2014.
[7] Ahsan MM, Siddique Z. Machine learning-based heart disease diagnosis: a systematic literature review. Artificial Intelligence in Medicine. 2022.
[8] Ibrahim I, Abdulazeez A. The role of machine learning algorithms for diagnosing diseases. Journal of Applied Science and Technology Trends. 2021; 2(1):10-9.
[9] Shaheen MY. Adoption of machine learning for medical diagnosis. ScienceOpen Preprints. 2021.
[10] Ahsan MM, Mahmud MA, Saha PK, Gupta KD, Siddique Z. Effect of data scaling methods on machine learning algorithms and model performance. Technologies. 2021; 9(3):1-17.
[11] Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. Journal of Big Data. 2019; 6(1):1-25.
[12] Osop H, Sahama T. Data-driven and practice-based evidence: design and development of efficient and effective clinical decision support system. In improving health management through clinical decision support systems 2016 (pp. 295-328). IGI Global.
[13] Bernell S, Howard SW. Use your words carefully: what is a chronic disease? Frontiers in Public Health. 2016.
[14] Philip R, Mathias M, KM DG. Evalation of relationship between markers of liver function and the onset of type 2 diabetes. Journal of Health and Allied Sciences NU. 2014; 4(2):90-3.
[15] Santos-gallego CG, Rosenson RS. Role of HDL in those with diabetes. Current Cardiology Reports. 2014; 16(9):1-4.
[16] https://www.astera.com/type/blog/pdf-scraping/. Accessed 20 September 2021.
[17] Blonce A, Filiol E, Frayssignes L. Portable document format (pdf) security analysis and malware threats. In presentations of Europe BlackHat 2008.
[18] Sumathi S, Esakkirajan S. Fundamentals of relational database management systems. Springer; 2007.
[19] Hashmi AM, Qayyum F, Afzal MT. Insights to the state-of-the-art PDF extraction techniques. IPSI Trans. Internet Res. 2020; 16(8):1-8.
[20] Ahmad R, Afzal MT, Qadir MA. Information extraction from PDF sources based on rule-based system using integrated formats. In semantic web evaluation challenge 2016 (pp. 293-308). Springer, Cham.
[21] Sateli B, Witte R. An automatic workflow for the formalization of scholarly articles' structural and semantic elements. In semantic web evaluation challenge 2016 (pp. 309-20). Springer, Cham.
[22] Klampfl S, Kern R. Reconstructing the logical structure of a scientific publication using machine learning. In semantic web evaluation challenge 2016 (pp. 255-68). Springer, Cham.
[23] Azimjonov J, Alikhanov J. Rule based metadata extraction framework from academic articles. arXiv preprint arXiv:1807.09009. 2018.
[24] Achilonu OJ, Singh E, Nimako G, Eijkemans RM, Musenge E. Rule-based information extraction from free-text pathology reports reveals trends in South African female breast cancer molecular subtypes and Ki67 expression. BioMed Research International. 2022.
[25] Mandal A, Bhattarai B, Kafle P, Khalid M, Jonnadula SK, Lamicchane J, et al. Elevated liver enzymes in patients with type 2 diabetes mellitus and non-alcoholic fatty liver disease. Cureus. 2018; 10(11).
[26] Bhowmik B, Siddiquee T, Mujumder A, Afsana F, Ahmed T, Mdala IA, et al. Serum lipid profile and its association with diabetes and prediabetes in a rural Bangladeshi population. International Journal of Environmental Research and Public Health. 2018; 15(9):1-12.
[27] Singh A, Dalal D, Malik AK, Chaudhary A. Deranged liver function tests in type 2 diabetes: a retrospective study. International Journal of Science and Healthcare Research. 2019; 4(3):27-31.
[28] Majid MA, Bashet MA, Moonajilin MS, Siddique M. A study on evaluating lipid profile of patients with diabetes mellitus. 2019.
[29] Shahwan MJ, Khattab AH, Khattab MH, Jairoun AA. Association between abnormal serum hepatic enzymes, lipid levels and glycemic control in patients with type 2 diabetes mellitus. Obesity Medicine. 2019.
[30] Islam S, Rahman S, Haque T, Sumon AH, Ahmed AM, Ali N. Prevalence of elevated liver enzymes and its association with type 2 diabetes: a cross-sectional study in Bangladeshi adults. Endocrinology, Diabetes & Metabolism. 2020; 3(2).
[31] Blomdahl J, Nasr P, Ekstedt M, Kechagias S. Moderate alcohol consumption is associated with advanced fibrosis in non-alcoholic fatty liver disease and shows a synergistic effect with type 2 diabetes mellitus. Metabolism. 2021.
[32] Tham YK, Jayawardana KS, Alshehry ZH, Giles C, Huynh K, Smith AA, et al. Novel lipid species for detecting and predicting atrial fibrillation in patients with type 2 diabetes. Diabetes. 2021; 70(1):255-61.
[33] Kosmalski M, Ziółkowska S, Czarny P, Szemraj J, Pietras T. The coexistence of nonalcoholic fatty liver disease and type 2 diabetes mellitus. Journal of Clinical Medicine. 2022; 11(5):1-24.
[34] Altaf I, Butt MA, Zaman M. Disease detection and prediction using the liver function test data: a review

of machine learning algorithms. In international conference on innovative computing and communications 2022 (pp. 785-800). Springer, Singapore.
[35] Godfrey KR. Correlation methods. Automatica. 1980; 16(5):527-34.
[36] Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. In noise reduction in speech processing 2009 (pp. 1-4). Springer, Berlin, Heidelberg.
[37] Sedgwick P. Spearman's rank correlation coefficient. BMJ. 2014.
[38] Abdi H. The kendall rank correlation coefficient. Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA. 2007:508-10.
[39] Aparicio M, Costa CJ. Data visualization. Communication Design Quarterly Review. 2015; 3(1):7-11.
[40] Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Computing and Applications. 2020; 32(24):18069-83.
[41] Wang Y, Han F, Zhu L, Deussen O, Chen B. Line graph or scatter plot? automatic selection of methods for visualizing trends in time series. IEEE Transactions on Visualization and Computer Graphics. 2017; 24(2):1141-54.
[42] Moon KW. Bar plot (I). In Learn ggplot2 Using Shiny App 2016 (pp. 111-20). Springer, Cham.
[43] Hicks SC, Okrah K, Paulson JN, Quackenbush J, Irizarry RA, Bravo HC. Smooth quantile normalization. Biostatistics. 2018; 19(2):185-98.
[44] Zheng A, Casari A. Feature engineering for machine learning: principles and techniques for data scientists. " O'Reilly Media, Inc."; 2018.
[45] Köpp C, Von MHJ, Breitner MH. Decision analytics with heatmap visualization for multi-step ensemble data. Business & Information Systems Engineering. 2014; 6(3):131-40.
[46] Friendly M. A brief history of the mosaic display. Journal of Computational and Graphical Statistics. 2002; 11(1):89-107.
[47] Friendly M. Graphical methods for categorical data. Proceedings of SAS SUGI. 1992; 17:1-7.
[48] Demšar J, Leban G, Zupan B. FreeViz-an intelligent multivariate visualization approach to explorative analysis of biomedical data. Journal of Biomedical Informatics. 2007; 40(6):661-71.
[49] Demsar J, Leban G, Zupan B. Freeviz-an intelligent visualization approach for class-labeled multidimensional data sets. Proceedings of IDAMAP. 2005; 1:13-8.

Ifra Altaf is a Research Scholar in the Department of Computer Sciences, University of Kashmir, Srinagar, J&K, India. She received her Master's degree in Computer Applications (MCA) and Bachelor's Degree in Computer Applications (BCA) from the University of Kashmir, J&K, India.
Email: hussainifra3@gmail.com

Muheet Ahmed Butt is a Scientist "D" in the Post Graduate Department of Computer Science, University of Kashmir, Srinagar, India. He received his PhD degree from the University of Kashmir, J&K, India and M.Tech in Communications and Information Technology from the National Institute of Technology [NIT] Srinagar. Additionally, he holds a Bachelor of Science in Computer Science Engineering from Bangalore University, India.
Email: ermuheet@gmail.com

Majid Zaman is Scientist "E" in the Directorate of Information Technology & Support System, University of Kashmir, J&K, India. He holds a PhD in Computer Science from the University of Kashmir, Srinagar, J&K, and an M.S. in Software Systems from the Birla Institute of Technology and Science (BITS), Pilani, India. In addition, he received a Bachelor's in Computer Science Engineering from BAMU, Mumbai, India.
Email: zamanmajid@gmail.com

Appendix I

S. No.	Abbreviation	Description
1	ALP	Alkaline Phosphatase
2	ANCOVA	Analysis of Covariance
3	AST	Aspartate Aminotransferase
4	CSV	Comma Separated Values
5	EHR	Electronic Health Record
6	EMR	Electronic Medical Record
7	ETL	Extract, Transform and Load
8	GGTP	Gamma Glutamyl Trans Peptidase
9	HbA1c	Glycated Haemoglobin
10	HDL	High Density Lipoproteins
11	IE	Information Extraction
12	LDL	Low Density Lipoproteins
13	LFT	Liver Function Test
14	LPP	Lipid Profile Panel
15	NLP	Natural Language Processing
16	PDF	Portable Document Format
17	SPSS	Statistical Package for Social Sciences
18	SQL	Structured Query Language
19	TC	Total Cholesterol
20	TG	Triglycerides
21	XML	Extensible Mark-up Language