Deep Learning-Based Semantic Diagnostic Framework for Big EHR Corpora

Sarah Shafqat [1,*], Qaisar Javaid [2], Majed Al-Saeed [3], and Hafiz Farooq Ahmad [3,*]

[1] Ambient Cloud, Islamabad, Pakistan
[2] Faculty of Computing and IT, International Islamic University, Islamabad, Pakistan
[3] Computer Science Department, College of Computer Sciences and Information Technology (CCSIT), King Faisal University, Al-Ahsa, Saudi Arabia
Email: ceo@DreamzSoft.org (S.S.); qaisar@iiu.edu.pk (Q.J.); alsaeed@kfu.edu.sa (M.A.S.); hfahmad@kfu.edu.sa (H.F.A.)
*Corresponding author

***Abstract*—Diagnoses if mishandled or left untreated may cause several other chronic illnesses as in the case of Diabetes Mellitus (DM). Researchers have explored endocrine domain for diagnoses of DM and its comorbidities using big data cloud analytics. It is seen that if DM patient condition gets worse it forms diseases like; breast cancer, arthritis, body pains, dementia, foot ulcers resulting in amputation, etc. Researchers initiated the design and development of unified medical corpora with real-time big Electronic Health Records (EHR) dataset of endocrine patients. The corpora are standardized using standard international medical nomenclature. It used International Classification Disease Diagnostic (ICD-10-CM) codes to label target diagnoses and is in compliance with Health Level Seven Fast Healthcare Interoperability Resources (HL7 FHIR) standard. Corpuses in the corpora with varying sizes and parameters is experimented on; Colab, Orange and RapidMiner open-source cloud frameworks. The performance of different models was observed using deep learning heuristics. The three models were proposed; i) Deep Neural Nets (DNN) Bi-directional Long Short Term Memory (Bi-LSTM) Sequential Model, ii) Louvain Mani-Hierarchical Fold Learning (LMHFL) and its optimized model; Fast-LMHFL, and iii) Custom Deep Multinomial Distribution Learning (DMDL) model enabled for semantics and temporal variables. DNN Bi-LSTM shows the efficacy of TensorFlow and deep neural networks. LMHFL gave diagnostic visual inferences with maximum accuracy of 0.727 and correlations up to 0.952 between patients' phenotypes with Laplace. DMDL was 100% accurate.**

***Keywords*—big data healthcare analytics, semantics, Named Entity Records (NER), tabular text mining, International Classification of Diseases (ICD-10-CM), endocrine comorbidity diseases, diagnostics**

I. Introduction

Vast data generation has led to information extraction for knowledge engineering in all fields as complex as healthcare. Authors started off with their research by thorough study of existing study on medical/clinical systems as illustrated in Fig. 1. Enabling medical data for information gain and automated decision support for clinical processes is highly researched upon while it keeps an eye on the issues that slows down the progression [1]. Although a great hype is felt for research in smart health field as we searched for a solution to diagnose endocrine patients suffering from Diabetes Mellitus (DM) and the coexisting diseases as referred to comorbidities. We came to know that a large Electronic Health Records (EHR) diagnostic data of patients gives a lot of room for improvement in performance of these smart health clinical systems.

There are several datasets in health informatics being maintained globally and foremost are; Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC) [2] and National Centre for Biotechnology Information (NCBI) [3]. MIMIC III version is the latest and it is accessible on request and attaining a course. These and similar health datasets are growing but limited in size, scope and its accessibility.

Therefore, as described in Fig. 1, to come-up with an interoperable dataset or corpora to be universally available for analysis over cloud, firstly, researchers contributed in the development of the unified medical corpora. It is based on unified clinical data model using Data Flow Diagrams (DFDs) and Entity Relationship Diagrams (ERDs) following Health Level Seven Fast Health Interoperable Resources (HL7 FHIR) standard and labeled with diagnostic codes for International Classification of Diseases (ICD-10-CM). These endocrine patients' corpora are of varying sizes, data types and complexities to test and validate any future analytical model. Researchers then contributed three proposed models that are mentioned here and in the published conference paper [4]. Further contributions involved authors use of open-source cloud platforms to explore the real-time endocrine diagnostic datasets extensively that ware provided in parts by Management Information System (MIS), Shifa International Hospital, Pakistan. The patient phenotypes not only include numeric fields like; age, but mostly it contains text fields; gender, clinical notes, practitioner comments, laboratory examination and results with target

Manuscript received January 6, 2025; revised January 13, 2025; accepted February 26, 2025; published May 22, 2025.

fields; diagnosis and its corresponding ICD-10-CM code with prescribed treatments; medicine and dosage, in some cases. Therefore, during the ad-hoc experimentation as mentioned in Fig. 1, researchers used several machine learning techniques and deep learning heuristics to formulate Big Data Analytics (BDA) for healthcare. Researchers gave emphasis on the presence of the text fields to embed semantic intelligence where appropriate using Natural Language Processing (NLP) techniques; Named Entity Records (NER) embedding and explored various other methods like; Fast.ai, etc.

The real-time endocrine datasets/corpuses as part of unified medical corpora with varying sizes and parameters is experimented mainly on; Colab, Orange and RapidMiner open-source cloud frameworks. The performance of different analytical models was observed using machine and deep learning heuristics. The three models were proposed; i) Deep Neural Nets Bidirectional Long Short Term Memory (DNN Bi-LSTM) Sequential Model, ii) Louvain Mani-Hierarchical Fold Learning (LMHFL) and its optimized model; Fast-LMHFL, and iii) Custom Deep Multinomial Distribution Learning (DMDL) model enabled for semantics and temporal variables. DNN Bi-LSTM Sequential Model validates the efficacy of using high performance frameworks like; tensorflow and deep neural networks to analyze textual big data. Diagnostic visual inferences gained from LMHFL, a hybrid deep learning algorithm on Orange show the associations in DM and other comorbidity diseases through rule mining having with maximum accuracy of 0.727 and correlations up to 0.952 between patients' phenotypes with Laplace. DMDL with 100% accuracy is found promising to come up with precisely accurate and speedy diagnosis.

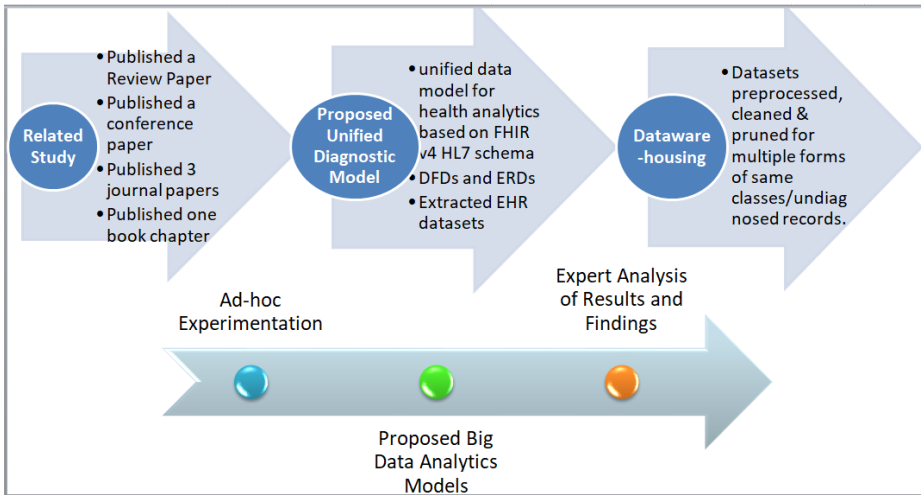

Fig. 1. Research methodological contributions projected in stages.

The paper is organized to start with introduction followed by related work. Section III is allocated to devise methodology with problem formulation and conceptualizing its solution with elaboration on corpora development and its value. Section IV elaborates experimentations and results at length. Sections V and VI are for discussion, conclusion, and future work.

II. Related Work

Big Heterogeneous Data in healthcare is being produced at fast rate which requires to be managed to extract and analyze information for clinical and administrative purposes. Researchers have carefully crafted their study in Table I and discussed the essence of it here. In clinical scenario it is analyzed for prediction, detection, diagnosis, and treatment of various diseases coming in from range of different platforms [5]. Specific to forming a medical diagnosis there are some most common risks and precautions associated. The complexity is well understood as the healthcare domain is vast and huge risk is associated with it. Therefore, the research is being carried out in various healthcare domains simultaneously by coordination of healthcare providers and technological community in support of World Health Organization (WHO) and international standards body for healthcare that is Health Insurance Portability and Accountability Act (HIPAA).

Kavakiotis *et al*. [6] established that diagnostic analytics are built on machine learning and data mining principles. Magrabi *et al*. [7] puts light on challenges that hover around using Artificial Intelligence (AI) in clinical decision support system or health information system and validating it. Researchers often ignore ethical issues while focusing on diagnostics for high accuracy and precision which is not recommended.

Reliance on training data limits the cognitive capabilities of system that may become reason for erroneous diagnosis. The wide spread of clinical data has yielded many algorithms and while selecting one poor fit of foundational data to a new setting may cause adherence issue. Another challenge is to keep machine learning algorithms up to date in knowledge. The concern, therefore, arises for generalization of an algorithm on all facets of clinical setting.

It is found necessary to include clinicians to judge the merits of algorithm to avoid conflict in its commercialization. Transparency of algorithm design may add cognition and engagement of patients and doctors.

Comprehensible Knowledge Model (CKMT) [8] is proposed in view of assisting clinicians the best recommended treatment plan for cancer through acquiring wealth of data available in clinical documents and EHRs. Data is refined into standard data format from various formats. Dataset of around 4000 head and neck cancer patients was taken that was distributed in approximately 14,000 documents. Classification and Regression Trees (CRT) algorithm were selected that produced 11 decision paths and yielded 11 rules reaching 69% accuracy level. This proposed methodology is adaptable to be applied on other domains like; for disease progression towards comorbidities of cardio, nervous system and gastro [9].

Predictive Analytical Decision Support System (PADSS) [10] was proposed for prediction [11] and identification of increasing disease outbreaks with response and confirmation. This was a healthcare cloud platform with Message Oriented Middleware (MOM) specified by Health Level Seven (HL7) platform having Fast Healthcare Interoperability Resources (FHIR) standard. The common high-performance tools used for big data analytics like; hadoop, mapreduce, azure, big query and jaql, are compared in Ref. [12], where it is seen that Amazon Cloud Platform is missing which is progressing at fast pace to include Machine Learning (ML) infrastructure for large-scale analytics.

In healthcare, decisions regarding disease diagnostics [13] through analytics for DM, heart, liver, cancer, skin, eye, neurological require clustering and classification using different tools to identify patterns in data stored as EHR or other formats in clinical practices taken by physicians [14]. Even though there are some public healthcare datasets available but still these are lacking in many ways [13]. Further, algorithms integrated into analytics come in various forms of machine learning and data mining techniques as shown in Table I. Mainly, in analytics, data mining and machine learning approaches are used. Segmentation, clustering, rule-based associations and classification concepts lies within the branch of data mining. When deciding if clustering or classification is done on labeled or unlabeled dataset, if analysis is done on runtime or prior stored data then this approach would be named as machine learning [15]. This approach is used for naming our learning data model for analytics as supervised (learning through labeled offline data), unsupervised (learning on data entered at runtime) or semi-supervised (using both known data and new unlabeled data that is adding to the storage space). Further machine learning approach transforms to deep learning when self-learning multilayer neural networks takes its place for optimization [16] by forming data hierarchy for unlabeled complex datasets to represent best features in data through exploration. Mainly, previous studies have used such techniques for facial-based analysis or image analysis [17, 18]. The proposed BDAs composed of data mining and advanced machine learning techniques assisted with Deep Learning (DL) mechanisms application over the data model is to be derived to integrate over SmartHealth cloud platform [19].

TABLE I. PREVIOUS STUDIES UNDERTAKEN FOR HEALTHCARE ANALYTICS

Cite	Proposed Technique/ System	Application	Algorithms/ Platforms	Testing/Validation Technique	Accuracy Level	Strength	Limitations
Afzal *et al*. [8]	Comprehensible Knowledge	generalized	Classification and Regression Tree (CRT) algorithm with 11 rules	Compared different results to select the right model	69% accurate	head/neck cancer treatment	Small dataset
Neto and Ferraz [10]	Cloud Based Predictive Analytical Decision Support System (PADSS)	Decision Support System (DSS)	PADSS built on top of PREVENT	2 simulations deployed on Google Cloud	HL7-FHIR	Real-time Big Data Analytics (BDA) Complex Event Process (CEP)	tuning of SQL queries required
Kale *et al*. [20]	BDA healthcare challenges/ promises	monitor, predict, IoT enabled, generate alerts	BDA healthcare framework	Out of 1,500 global people, 72% unaware of BDA. 80% hasn't seen it. 20% stated it as ineffective	Detailed review suggests smart hospitals are highly desirable	80% of 1500 believed it. 93% agreed for treatment in smart hospital	Unstructured clinical data, confidentiality, partition & discontinuity
Reddy *et al*. [21]	Azure Machine Learning (ML) platform	design/develop robust binary/multiclass models	binary and multilevel forest decision making	Validated UNSW NB-16 real-time cyber-attacks HDFS distributed storage	99.2% in binary, 99%, 94.49%, 91.79%, & 90.9%, multi-level forest model	train 176,341 network instances in 6 s	hard to manage variable parameters transformed to integers
Howard and Gugger [22]	Fastai Deep Learning (DL) library	build DL approaches from low-level components	Uses TensorFlow over Pytorch GPU library	Library is widely used	State-of-the-art	Standard DL allows optimized few lines of code; novel 2-way callback optimizer	fastai always has room for development
Raghu and Schmidt [23]	Survey of DL, Deep Neural Nets (DNNs), Tensorflow/kera	Scientific Discovery	Open-source BDA resources	validates DL flow	Review	fast DL for images/language	get big data, less data & right model

Zhu *et al*. [24]	clinical concept extraction NER	clinical concept embedding	Tokenize using ELMo ensemble with Bi-LSTM-CRF, NLTK	Validated on SNOMED-CT ontology extracted from free text	3.4% more accurate	NER embedding	Un-predictive
Kaggal *et al*. [25] Foley *et al*. [26]	Learning Healthcare System	web-based application AskMayoExpert (AME)	implemented MEA NLP BDA on Unified Data Platform (UDP)	3 pilot Care Process Models (CPMs); hyperlipidemia, atrial fibrillation & Congestive Heart Failure (CHF)	standard	HL7	Confined in hospital premises
Silva *et al*. [27]	Healthcare Diabetes BDA	Clinical DSS	dimensional models &Apache Mahout Hadoop	Criteria to judge sparse complex health big data 6 Vs	DM Screen Complications Research Initiative (DiScRi)	Diagnostic data stages give opportunities/ challenges	Complex health data slows BDA
Jankowski [28]	Clinical Medicine (Psychological)	IncNet Kalman Filter bi-central multi-dimension	Minnesota Multiphasic Personality Inventory (MMPI)	Not mentioned	IncNet 99.23% more accurate than C4.5 & FSM	optimized IncNet & FSM	small data.
Grochowski and Jankowski [29]	DM, heart, breast, iris, skin cancer etc.	instance selection algorithms comparison	UCI ML	KNN, SVM, SSV, NRBF, FSM, IncNet	RMHC, MC1, LVQ, DROP2-4 & DEL found effective	stabilized learning	limited datasets
Kahramanli and Allahverdi [30]	Endocrine	Diagnostic Hybrid System (ANN & FNN)	Pima Indians Diabetes	Matlab 7.0.0	Outperformed IncNet with other	Outperformed	limited dataset
Nalluri *et al*. [31]	DM, Liver, Breast Cancer, Hepatitis, etc.	Diagnostic Hybrid Intelligent System	11 datasets from UCI (Pima Indian Diabetics & BUPA Liver Disease)	Weka	Optimized Firefly with MLP & SVM	Optimized hybrid analytics	limited Dataset

III. Methodology

To finalize our problem, we explored healthcare analytics domain and formed a thematic taxonomy [13] of its application in various health domains. Diagnostics of endocrine diseases held our attention. It started-off with the initiation to build corpora comprising of standardized real-time Big EHR Data for generalizability and interoperability. We devised our research methodology by formulating the problem from the limitations found in current research and followed the research interests in the health informatics community. The gaps were found and required to handle:

(1) Unavailability of standardized diagnostic system:
   a). Traditionally, emergencies are not predictable in advance and mostly occur unexpected and as surprise for patient, his/her relatives and doctors.
   b). In case of chronic diseases even the expert physicians sometimes find themselves ignorant with the latest research or trend being adopted internationally in diagnostic measure due to lack of connection or resources.
   c). Patients often find it difficult to connect to the relevant doctor or expert in the field.

(2) Modeling healthcare big data:
   a). Conversion of unstructured or semi-structured EHR big data [18] coming in various formats into structured machine-readable diagnostic data model with best feature set [6] using medical ontology based on ICD-10-CM Diagnostic coding and HL7 FHIR v4 [32] standards is a challenge.
   b). Maintaining, mining and selecting the right feature set from vast EHR data specific to disease and diagnosis is not easy [6].

(3) Designing analytics to process endocrine big data:
   a). To find an optimum big data healthcare analytics technique for diagnosis of diabetes and its comorbidities [33] as recent research [6, 33] has evaluated several diagnostic analytics. The challenge, therefore, is to link data analytics with diagnosis and appropriate decision-making in drug management to avoid risks of forming comorbidities.
   b). Making a generalized analytical algorithm with transparency for building trust and conformance for patients and doctors engaged at different clinical settings is an issue [7].
   c). An interoperable intelligent diagnostics system is therefore required which would take lots of memory on run-time. Currently, most analytics are synchronous in nature and do not offer parallel processing asynchronously.

A. The Proposed Approach

It is understood that the clinical medical corpora would be required of endocrine patients to be fed into big data analytics for diagnosis and prognosis.

Researchers decided to initiate the design and development of the unified medical corpora in compliance with international standards for generalization and interoperability over cloud platforms.

Optimum big data healthcare analytics for clinical purposes are then designed, developed and experimented for validation using the corpora. The unified endocrine corpora fetched from MIS, Shifa International Hospital, Pakistan, when fed into the proposed analytics eases the process of diagnosing DM and the comorbid diseases with in-time prognosis specific to patient's medical profile. The proposed healthcare big data analytics were prepared and validated on multiple open-source cloud platforms.

1) Step/module one: proposed unified clinical data model

The unified clinical data model is designed with detailed consultation and approval by the technical staff at MIS department of Shifa International Hospital before fetching the datasets to transform into desired corpora; DM_Comorbid_EHR_ICD10.

Researchers also kept the data model features follow international standards so that it is adopted and used globally over interoperable cloud for smart clinical decision support system. The proposed clinical data model, therefore, is in compliance with the HL7 FHIR v4.0 schema and has the target diagnostic classes labeled with ICD-10-CM standard codes manually for precision.

2) Step/module two: proposed unified endocrine corpora

The endocrine EHR datasets were fetched from Shifa International Hospital, Pakistan, against the unified clinical data model based on HL7 FHIR v4.0 standard for generalizability and interoperability. These variable sized datasets went through rigorous data warehousing steps and were put into corpora; DM_Comorbid_EHR_ICD10, having;Corpus100_DM_pts_2844,Corpus100_DM_pts_9304, and Corpus14407_DM_pts_33185.

3) Step/Module three: proposed experimental design on open-source cloud platforms

We experimented with three meta-datasets of endocrine patients with varying sizes and parameters to train analytics on chosen open-source analytical cloud platforms; i. Gradient Paperspace, ii. Google Colab, and iii. auto ML platforms like; a) Tableau, b) Qlik Sense. c) Orange and, d) RapidMiner.

4) Step/module four: proposed healthcare big data cloud analytics

Performance of the proposed big data analytical models; i) DNN Bi-LSTM Sequential Model, ii) LMHFL and the optimum model which is based on semantics; Fast-LMHFL, were evaluated, and iii) Custom DMDL model enabled for semantics and temporal variables was further evaluated and validated for optimal explainable results and maximum accuracy discussed by researchers in detail in [34–36].

B. Value of the Data

Unified clinical data model [1] in compliance with HL7 FHIR standard schema is built to extract input data to excel sheets to build diagnostics for DM and its comorbidities over the cloud.

This proposed model would prove to be very significant if adopted and extended as per the needs by other researchers in future to build smart clinical decision support systems and beyond.

The resource in our case is EHRs that are stored and managed by MIS departments in hospitals. We extracted this dataset against the entity tables in the proposed data model [1]. Records of 100 to 14,407 DM patients with comorbidities are kept in the dataset in parts provided by MIS department, Shifa International Hospital, Pakistan. The normalized datasets that were extracted went through rigorous data warehousing techniques. The data tables were de-normalized to form three flat tabular excel datasheets after extensive data wrangling, preprocessing and cleaning.

One unified corpora was formed; DM Comorbid EHR ICD10, with a naming convention; DiseaseName_Comorbid_EHR_StandardLabelConvention. It depicts the corpora has EHRs of endocrine patients who suffered DM with comorbidity diseases labeled with standard diagnostic codes set by International Classification of Diseases (ICD-10). The 3 Corpuses in it are further named as per the number of patients and their records or instances.

These corpuses are of different sizes, data types and complexities to accurately validate the big data healthcare analytics over the cloud platforms.

C. Datasets Elaborated for Experimental Setup

The unified medical corpora; DM_Comorbid_EHR_ICD10, is prepared from real-time patients' big EHR data provided by MIS, Shifa International Hospital, Islamabad, Pakistan. Excel sheets made from ERDs were imported into access database in normalized form.

These were then de-normalized into three flat datasheets that were named based on the number of diagnostic records from sets of 100 to 14,407 patients; Corpus100_DM_pts_2844, Corpus100_DM_pts_9304, and Corpus14407_DM_pts_33185.

Our focus was the diagnosis and prognosis of DM its comorbidities labeled with ICD-10 codes. Therefore, our corpora are named DM_Comorbid_EHR_ICD10.

The elements of DM Comorbid EHR ICD10 corpora are defined in Table II.

TABLE II. CORPORA EXPANDED WITH DETAILS OF THREE CORPUSES WITHIN

Corpus	No. of Patients	No. of Records (in raw form)	No. of Records after cleaning
Corpus100_DM_pts_2844	100	3,650	2,844
Corpus100_DM_pts_9304	100	15,696	9,304
Corpus14407_DM_pts_33185	14,407	87,803	33,185

Corpus100_DM_pts_2844 consists of 100 DM patients having four comorbidity diseases. The 100 patients corpus size gradually increases and becomes complex.

Corpus100_DM_pts_9304 has 65 classes of diagnostic labeled diseases with DM as primary disease. Multiple records are there for each patient diagnosed with DM and several other comorbid diseases in parallel.

Corpus14407_DM_pts_33185, is comparatively a larger corpus of 14,407 endocrine patients but these DM patients are diagnosed with only 32 labeled classes including DM.

There are total of 16 normalized entities. Main entities among all are; Patients, Tests, Medication, Allergies, and DiagnosticReport. Most relevant features are selected from the same entities to train our machine-learning algorithms for diagnosis as in Figs. 2–5. Fig. 2 shows that features "Test" and "Result" have significance in the dataset using "info- gain" and other relevant measures.

Ranks

	#	Info. gain	Gain ratio	Gini	ANOVA	χ^2	ReliefF	FCBF
Test	109.0	1.3001084718887959	0.5273190414719929	nan	nan	745.852492516492	0.08673464034763481	nan
Result		0.9296484204977133	0.4649114685530713	0.03418205861949364	nan	60.938022616454	0.015907216288731624	0.39767383095030634
PatientID	44.0	0.7588336225668895	0.1407249542397763	nan	nan	95.10926014701191	-0.10726301178489679	nan

Fig. 2. Features evaluation matrix.

Out[1]:

OLS Regression Results

Dep. Variable:	DiseaseID	R-squared:	0.010
Model:	OLS	Adj. R-squared:	0.008
Method:	Least Squares	F-statistic:	4.491
Date:	Thu, 19 Sep 2019	Prob (F-statistic):	0.0115
Time:	09:39:09	Log-Likelihood:	-3728.2
No. Observations:	884	AIC:	7462.
Df Residuals:	881	BIC:	7477.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	28.9036	1.277	22.641	0.000	26.398	31.409
C(HbA1c)[T.HbA1c < 5.7%]	21.0964	16.497	1.279	0.201	-11.282	53.475
C(HbA1c)[T.HbA1c >= 6.5%]	-3.7042	1.417	-2.615	0.009	-6.485	-0.924

Omnibus:	3407.859	Durbin-Watson:	0.404
Prob(Omnibus):	0.000	Jarque-Bera (JB):	94.981
Skew:	0.448	Prob(JB):	2.37e-21
Kurtosis:	1.668	Cond. No.	39.1

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fig. 3. Examination → test = diagnosis: (i) HbA1c ≤ 5.7% = considered normal, (ii) 5.7% < HbA1c < 6.4% = pre-diabetic or diabetic, (iii) HbA1c ≥ 6.5% = diabetic.

Out[4]:

OLS Regression Results

Dep. Variable:	DiseaseID	R-squared:	0.058
Model:	OLS	Adj. R-squared:	0.034
Method:	Least Squares	F-statistic:	2.462
Date:	Thu, 19 Sep 2019	Prob (F-statistic):	0.0917
Time:	09:47:51	Log-Likelihood:	-355.96
No. Observations:	83	AIC:	717.9
Df Residuals:	80	BIC:	725.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	22.8571	6.789	3.367	0.001	9.348	36.367
C(GF)[T.Normal]	-0.8571	9.600	-0.089	0.929	-19.963	18.248
C(GF)[T.Prediabetes]	11.2443	7.125	1.578	0.118	-2.934	25.423

Omnibus:	133.218	Durbin-Watson:	0.981
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8.760
Skew:	-0.264	Prob(JB):	0.0125
Kurtosis:	1.499	Cond. No.	8.46

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fig. 4. GF test is known to monitor diabetes patient. (i) GF < 70 == Low, (ii) 70 ≤ GF ≤ 99 == normal, (iii) 100 ≤ GF ≤ 125 == Pre-diabetes, (iv) GF ≥ 126 == diabetes.

Out[3]:

OLS Regression Results

Dep. Variable:	DiseaseID	R-squared:	0.002
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	1.292
Date:	Thu, 19 Sep 2019	Prob (F-statistic):	0.275
Time:	09:45:42	Log-Likelihood:	-6388.7
No. Observations:	1441	AIC:	1.278e+04
Df Residuals:	1438	BIC:	1.280e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	30.1655	1.186	25.439	0.000	27.839	32.492
C(TSH)[T.TSH < 0.4]	1.5074	2.301	0.655	0.513	-3.007	6.022
C(TSH)[T.TSH > 4]	-1.3573	1.344	-1.010	0.313	-3.994	1.280

Omnibus:	5698.717	Durbin-Watson:	0.181
Prob(Omnibus):	0.000	Jarque-Bera (JB):	213.776
Skew:	-0.088	Prob(JB):	3.80e-47
Kurtosis:	1.121	Cond. No.	5.97

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fig. 5. TSH tests thyroid: TSH < 0.4 mU/L = hyperactive thyroid, 0.4 ≤ TSH ≤ 4 == normal, TSH > 4 = underactive thyroid.

Out[4]:

OLS Regression Results

Dep. Variable:	DiseaseID	R-squared:	0.018
Model:	OLS	Adj. R-squared:	0.015
Method:	Least Squares	F-statistic:	6.597
Date:	Thu, 19 Sep 2019	Prob (F-statistic):	0.0106
Time:	12:32:15	Log-Likelihood:	-1604.5
No. Observations:	370	AIC:	3213.
Df Residuals:	368	BIC:	3221.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	30.9660	1.210	25.598	0.000	28.587	33.345
C(LDL)[T.Normal]	-5.1437	2.003	-2.568	0.011	-9.082	-1.206

Omnibus:	2359.070	Durbin-Watson:	0.667
Prob(Omnibus):	0.000	Jarque-Bera (JB):	42.277
Skew:	-0.044	Prob(JB):	6.60e-10
Kurtosis:	1.346	Cond. No.	2.42

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fig. 6. Bad cholesterol leads to heart disease: (i) LDL ≤ 100 = normal, (ii) LDL >100 = Bad, (iii) higher LDL > 190 raise chances of heart disease.

In Figs. 3–6, there are some statistical terms gained, from running Ordinary Least Square (OLS) statistical accuracy measure, that define the co-relation between

different tests with the diagnosed disease. OLS R-squared between 0 to 100% would tell if the data fits the model. Large value of F-Statistic would represent significance of values as for Diagnosis with HbA1c, Glucose Fasting (GF), Thyroid-Stimulating Hormone (TSH) or Low-Density Lipoprotein (LDL) in Figs. 3–5. Durbin-Watson below 2 would indicate positive correlation. p-value near "0" would refer to significance of results. Prob(Omnibus) closer to 0 would mean the model outperforms. Higher kurtosis would refer to a tighter cluster with fewer outliers.

Some important features selected as inputs are demonstrated in Tables III and IV. Each patient in flat de-normalized table has multiple diagnoses based on multiple examinations and tests done in accordance with the patient's condition/symptoms. Further accuracy in diagnoses maybe achieved based on previous disease suffered and medications that a patient has taken for it.

TABLE III. CORPUSES HAVE SOME IMPORTANT FEATURES IN TEXT AND NUMERIC FORM

PatientID	Age	Gender	VAN	Note	Examination	Test	Result	PC	ICD-10-CM	Diagnosed
62196	55	Female	180102cgf7	T2 Diabetes25 years humulin 70/30 20,20	Alk Phosphatase	Alkaline Phosphatase	100	Loss apetite nausea, Polyurea loss of conciuos ness and dizziness	24.92	HORMONAL
62196	55	Female	180102cgf7	T2 Diabetes25 years humulin 70/30 20,20	Bicarbonate Serum	Bicarbonate Serum	20	Loss apetite nausea, Polyurea loss of conciuos ness and dizziness	24.92	HORMONAL

TABLE IV. PC AND NOTE FIELDS IN CORPUSES WHEN EXPANDED CONTAINS CONDITION AND MEDICATION INFORMATION

PatientID	Age	Gender	VAN	Note	Examination	Test	Result	PC	Condition	Medication	Disease	ICD-10-CM	Diagnosed
102579	79	Female	180407jVh1	c/o dry mouth ankle swelling on Humulin R as per sliding scale	CBC Diff Profile (CS11)	MCHC	33.3	Known metastatic Carcinoma of ovary started on solumedrol and then deltacortil BG 400–500	dry mouth, ankle swelling	Humulin, solumedrol, deltacortil BG 400–500	metastatic Carcinoma of ovary	E09.9	Steroid Induced Diabetes

IV. EXPERIMENTATION, ANALYSIS, AND RESULTS

Exhaustive experimentation was run on the real-time unified endocrine datasets using data wrangling and warehousing techniques. The finalized endocrine corpora were then used for diagnosis of DM and the comorbidity diseases.

In result, we were able to propose big data healthcare analytical models with optimum results for multiple diagnoses of endocrine patients.

A. Qlik Sense or Tableau

Such platforms were used for preliminary exploration of data model and the important features. These were run on raw and finalized datasets in parallel as mentioned in Table II; Corpus100_DM_pts_2844 in raw form with 3650 records, Corpus100_DM_pts_9304 with 15,696 and Corpus14407_DM_pts_33185 records were inputted to get an idea of statistical inference between features as demonstrated in Figs. 7–24.

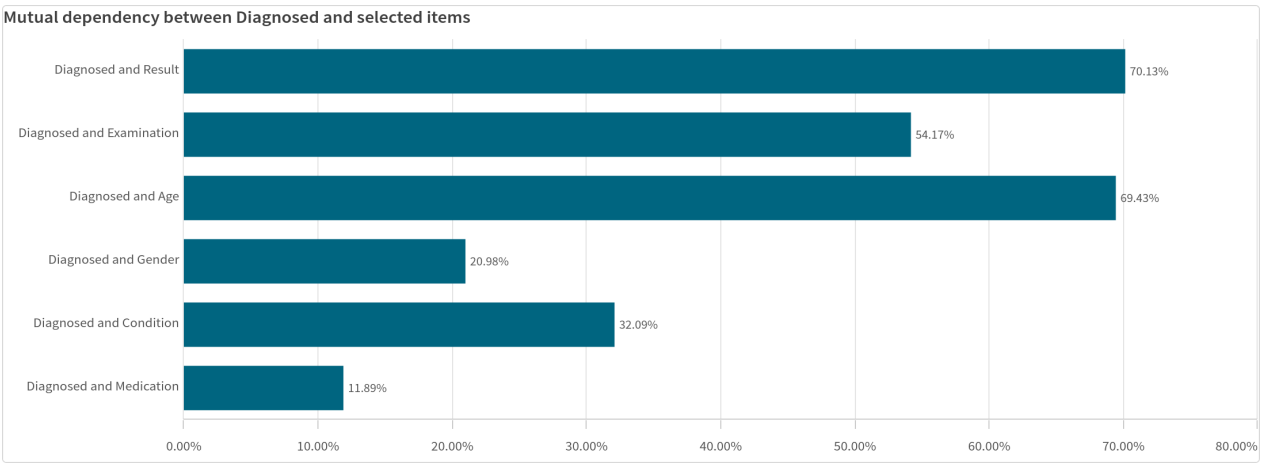

Fig. 7. Dependency of diagnosed disease on related feature sets (fields) in Corpus100_DM_pts_2844.

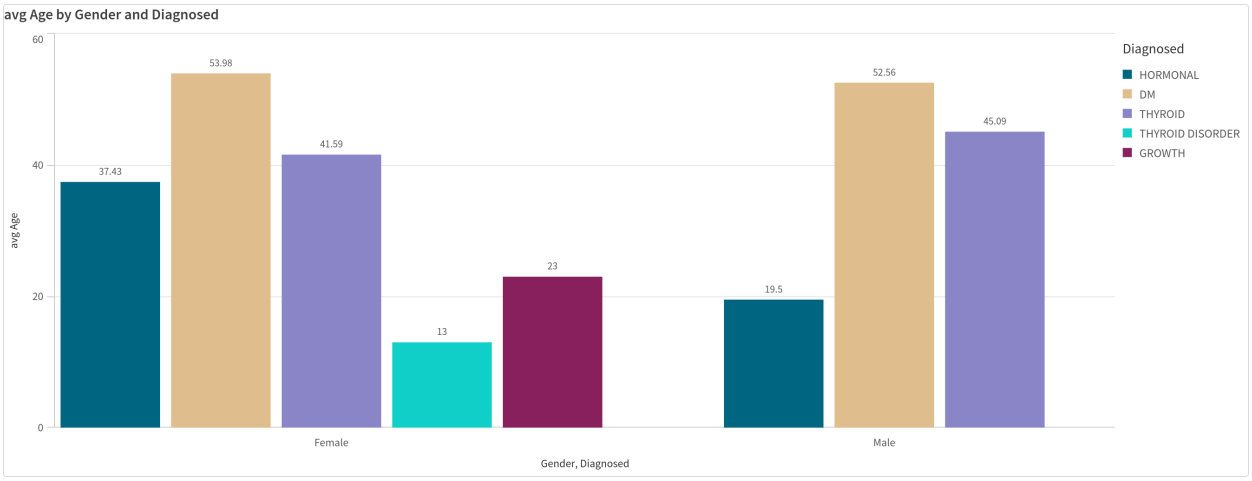

Fig. 8. Females have high ratio of forming endocrine diseases in Corpus100_DM_pts_2844.

In Fig. 7, target class Diagnosed, in the corpus with 2844 records, is shown to depend mostly on features; Lab Result, Age, Lab Examination Recommended, patient condition or symptoms, gender and medication, respectively.

Fig. 8 shows five Diagnosed classes in association with patients' Gender. DM is primary Diagnosed class and others are related co-existing thyroid and hormonal Diagnosed classes. Hormonal, thyroid or growth disorders or diseases is shown exceeding in females. On the other hand, males also show hormonal and thyroid issues with DM as primary disease.

Fig. 9 shows more proportion of females between ages of 20 to 60, suffering from endocrine diseases where males are older than 40 and are relatively less in number than women.

Endocrine diseases are seen more prevalent in women who start showing symptoms at much earlier age than men as shown in Fig. 10.

Among endocrine diseases, in corpus with 2844 records of 100 patients, it is seen that hormonal issues start to surface at much earlier age, DM is formed in late 20 s or 30 s where thyroid emerges in late 40 s (Fig. 11).

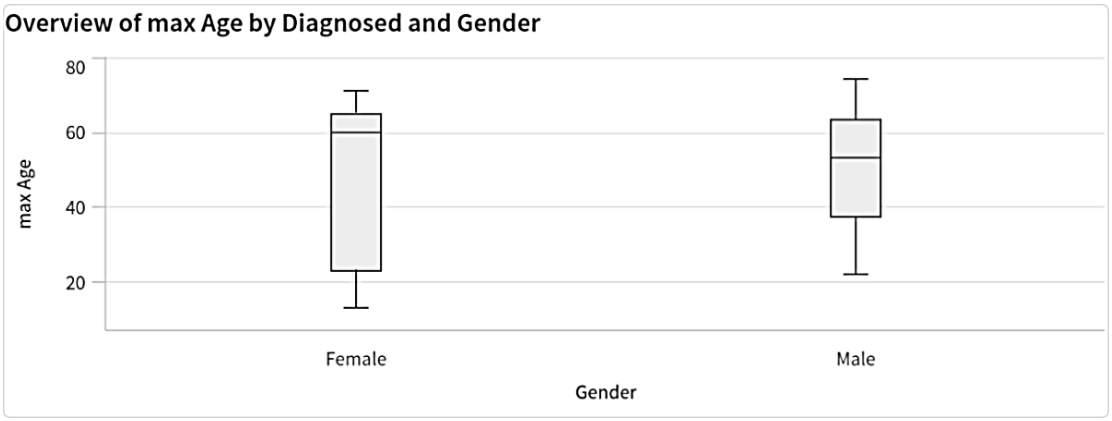

Fig. 9. Females between 20 to 60 and males between 40s to 60s are suffering from endocrine diseases.

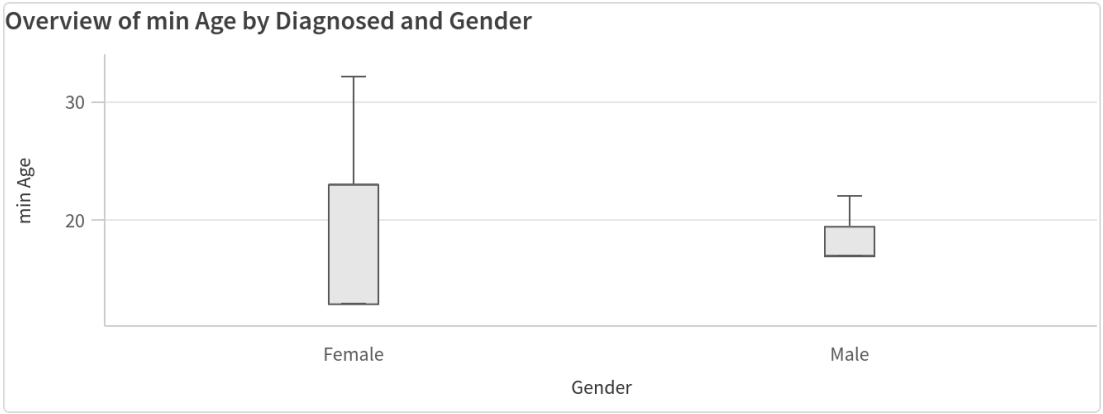

Fig. 10. Females are prone to forming endocrine diseases at much earlier age than men.

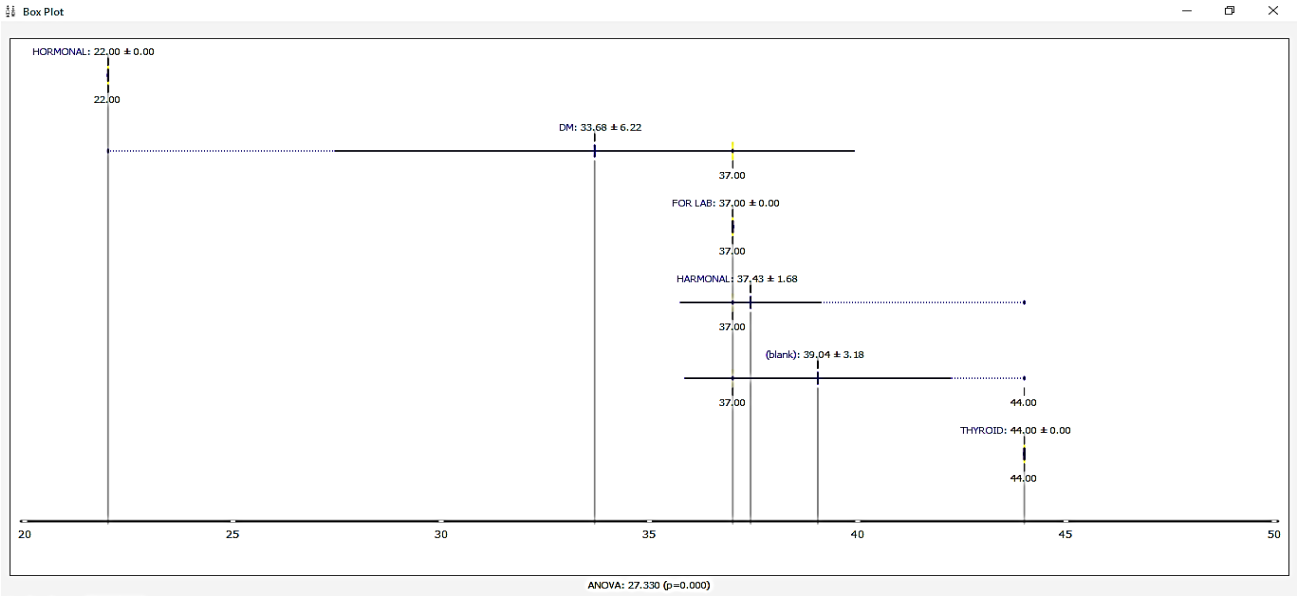

Fig. 11. Box plot shows patients with hormonal diabetes formed in early 20s, DM in 30s and thyroid in 40s.

Corpus having 9304 records of same 100 patients shows a larger set of endocrine or other coexisting diseases with DM. In Fig. 12, the diagnosis for breast cancer is seen prevalent where tests for White Blood Cells (WBC) equals 1300. Where in, Parathyroidectomy, WBC counts get as low as 940. Normal WBC count in males is 5000 to 10,000 and in females it is 4500 to 11,000 at birth. Patients with thyroid issues are seen deficient in vitamin B12 where normal borderline levels are 200 and 300 pg/ml. In Figs. 12 and 13, we also see same diagnosis with multiple labels for differences in text format or spellings with reflect data impurity that was corrected in parallel.

Diagnosis	Test		
	Urine Micro Albumin	Vitamin B12	WBC Total
CA BREAST	-	-	1300
CA BREAST	-	-	1300
CA breast	-	-	1300
Ca Breast	-	-	1300
ca breast	-	-	1300
F/Up Thyroidectomy and Parathyroidectomy	-	-	940
THYROID	< 5	32.1	-
thyroid	-	-	-
THYROID DISORDER	-	72.2	-
Thyroiditis	-	-	-
THYROIUD	-	-	-
total thyroidectomy	10	-	-

Fig. 12. Breast cancer is significantly apparent with WBC count = 1300 in Corpus100_DM_pts_9304.

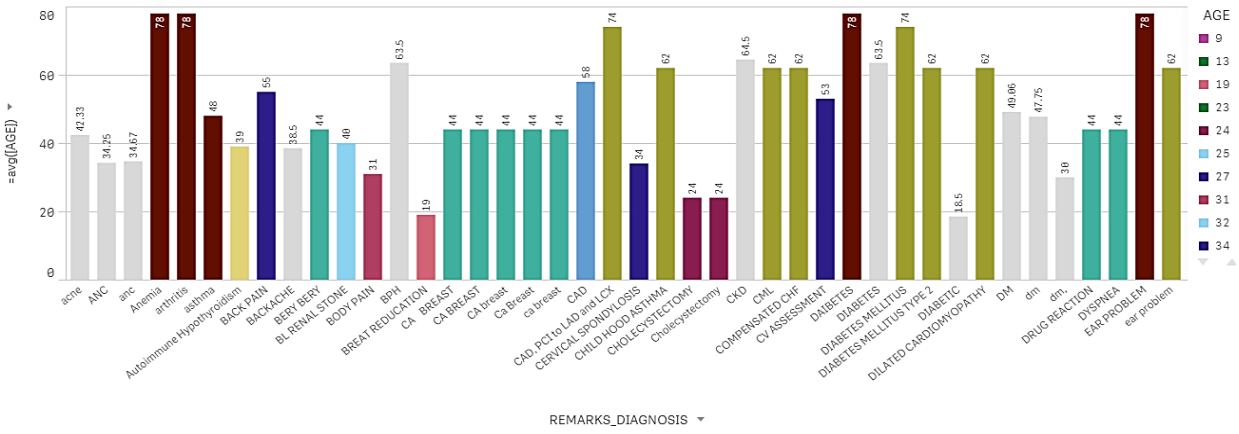

Fig. 13. View of raw diagnostic data with multiple labels for same diagnosis as “DIABETES”, “DIABETIC”, “dm”, and “DM” in Corpus100_DM_pts_9304.

In Fig. 14, the corpus is explored for lab examinations where a single examination is seen to have multiple recommended tests to reach specific diagnoses.

Tree or Heat Map is shown in Fig. 15, where for DM, hormonal and thyroid diseases, different lab examinations are seen recommended with respect to average age of patient.

Tree or Heat Map is shown in Fig. 16, where different lab examinations are seen recommended with respect to average age and gender of patient.

Above preliminary visualizations shown helped to understand the features in an elaborated way.

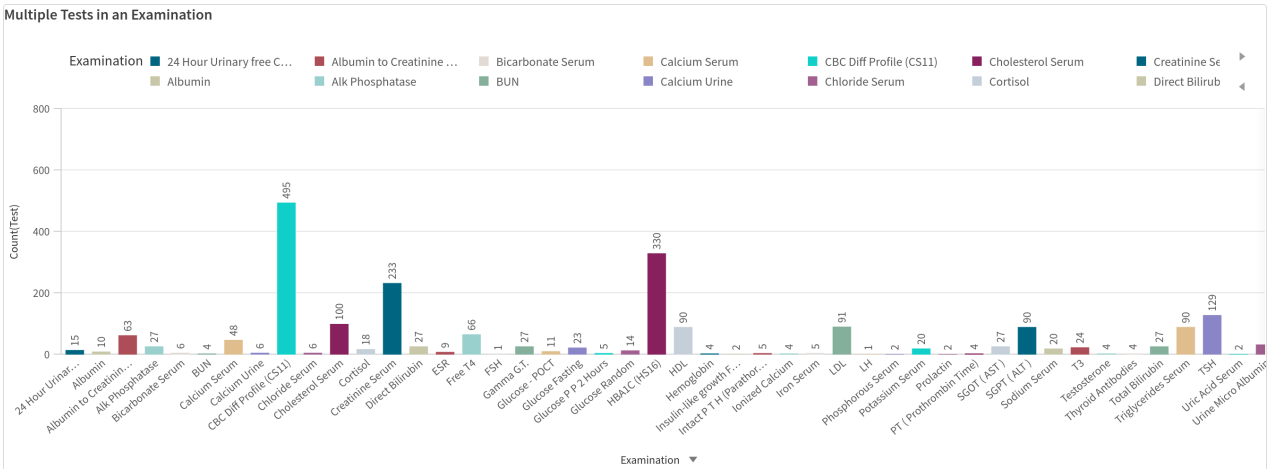

Fig. 14. One clinical examination has multiple tests for endocrine patients.

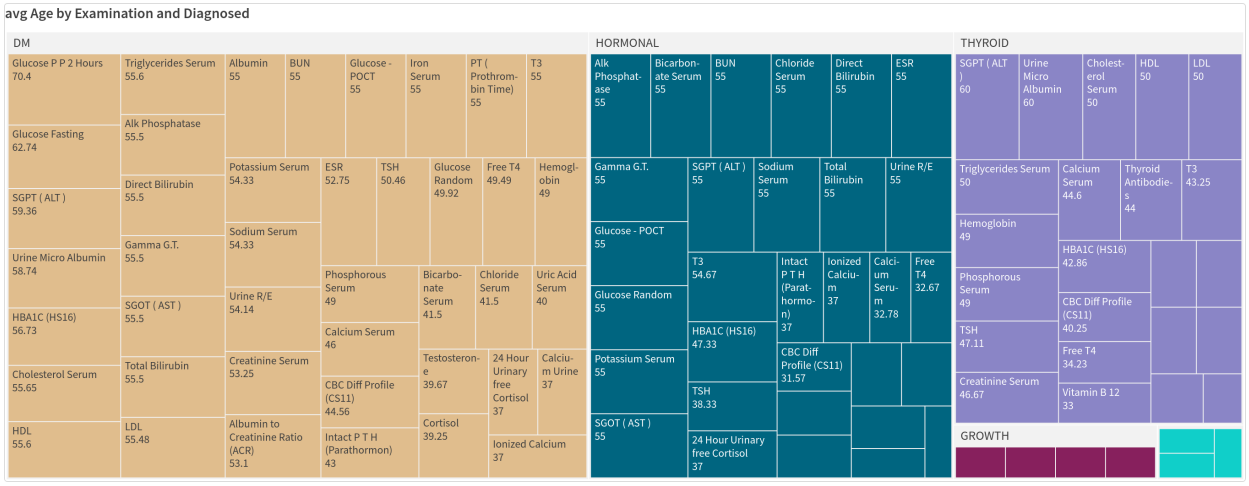

Fig. 15. The examinations recommended for patients (average age is shown) with DM and other coexisting diseases in Corpus100_DM_pts_2844.

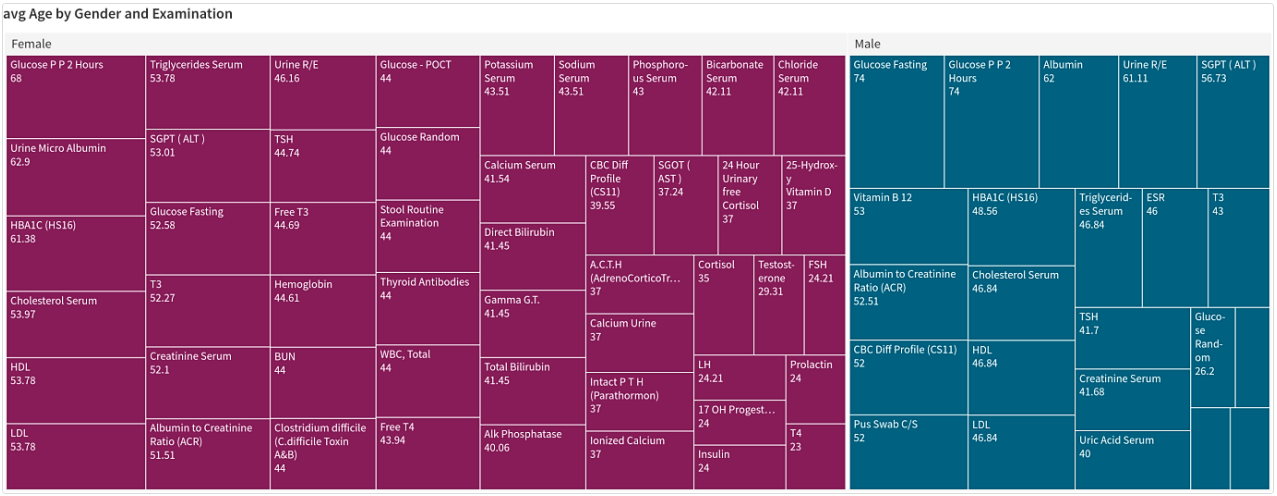

Fig. 16. Corpus100_DM_pts_9304 in pure form shows females and males with respect to average age undergo various examinations for endocrine diseases.

B. Gradient Paperspace and Google Colab

Such open-source cloud platforms were used to build analytical models for diagnostics.

Firstly, the traditional algorithms for AI and ML were tested employing Natural Language Processing (NLP) and NER methods to enable semantic learning in the proposed models. Manual and automated annotation techniques were explored and tested. HunFlair [25] based on NCBI disease corpus was evaluated that was found limited to label correct diagnosis for DM and its comorbidities. Therefore, we used spacy for manual annotation and to label our corpus and made a model.pt that was run in HunFlair. This transfer learning technique worked to correctly label the exams, tests, diagnosis and medicine to the corresponding conditions of the patients. The HunFlair transfer learning model would improve when we are able to annotate the whole corpora for the problem at hand to diagnose DM and its comorbid diseases.

Still we were able to distinguish the important diagnostic labels and NER embeddings from the PC and Notes fields and the features in our corpuses were expanded for better diagnosis. DNN Bi-LSTM Sequential model [34] on DNNs with Bi-LSTM Language using TensorFlow Keras which was found quite efficient in diagnosing DM and its comorbidities. Accuracy of 0.9 was achieved for comparatively bigger dataset with maximum input features. It was limiting as not all diseases were diagnosed but justified the reason for authors to explore deep learning heuristics as in [35]. The proposed analytical model was run on three platforms; Anaconda, Paperspace and Google Colab. The difference in terms of speed and performance was visible showed in Fig. 10 [34].

C. Orange Framework

It was used to design, simulate and validate Hybrid Deep learning heuristics named as LMHFL. It is found promising with clarity of diagnostic results through explainable visual inferences in form of clusters and graphs. LMHFL results in expert level visualizations with visible associations in features that form diagnosis of DM and correlated comorbidities. Diagnostic results gathered show the relation among DM and other comorbid diseases through rule mining having maximum accuracy of 0.7 and correlations up to 0.989 between patients' phenotypes with Laplace. Rules induced for single DM patient with comorbidity diseases gave the accuracy up to 0.9. These results were then compared to the results generated on textual tabular dataset run in Fast-LMHFL, an optimized form of analytical model [36]. Fast-LMHFL promised better results when run on single male and female patients of ages 73 and 44 respectively. Text-tabular Data posed a limitation to run Fast-LMHFL on the whole dataset as the fields; Note and PC in dataset were mostly missing.

Fig. 17 shows multiple endocrine diagnoses in a corpus of 100 patients with 2844 records where we see how each diagnosis relates and associates with the other in respect to age and patients count. Mostly patients diagnosed with DM, hormonal and thyroid are found in the age brackets of 34 and 40 with some exceptions.

In Fig. 18 we see diagnoses in result of different lab examinations. There is large ratio of females with some symptoms and conditions shown in clinical notes that form different related diagnoses. There are some patients highlighted to show the lab results for recommended tests.

D. RapidMiner

It was chosen to design, built and simulate custom auto ML diagnostics; Custom DMDL model [36] that showed 100% accuracy when Corpus100_DM_pts_2844 and Corpus14407_DM_pts_33185 were validated using decision tree during cross validation and split validation (Fig. 19). DMDL trained on complex Corpus100_DM_pts_9304 gave 97.3% accuracy due to the complexity it has with maximum number of classes above 65 and single patient has multiple diagnostic records for DM and its comorbidity diseases that a patient suffers from. Still the accuracy is seen to improve with multiple runs. Later in Fig. 20, the gender ratio of patients is shown where female patients exceed.

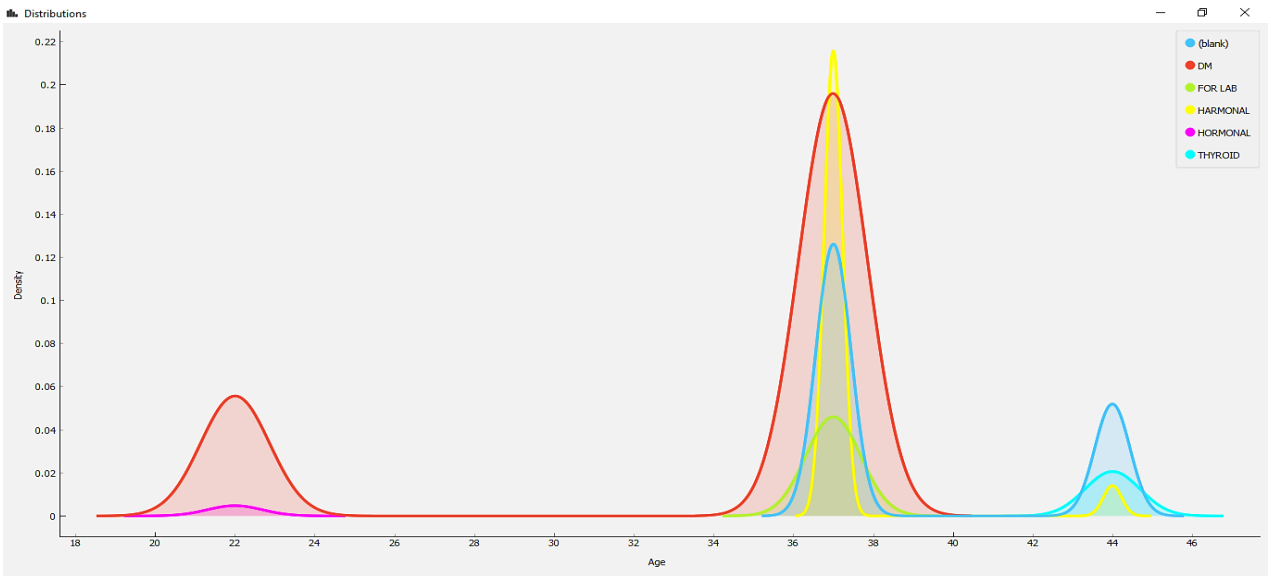

Fig. 17. DM ratio with age (found in early 20 s or late 30 s) in Corpus100_DM_pts_2844.

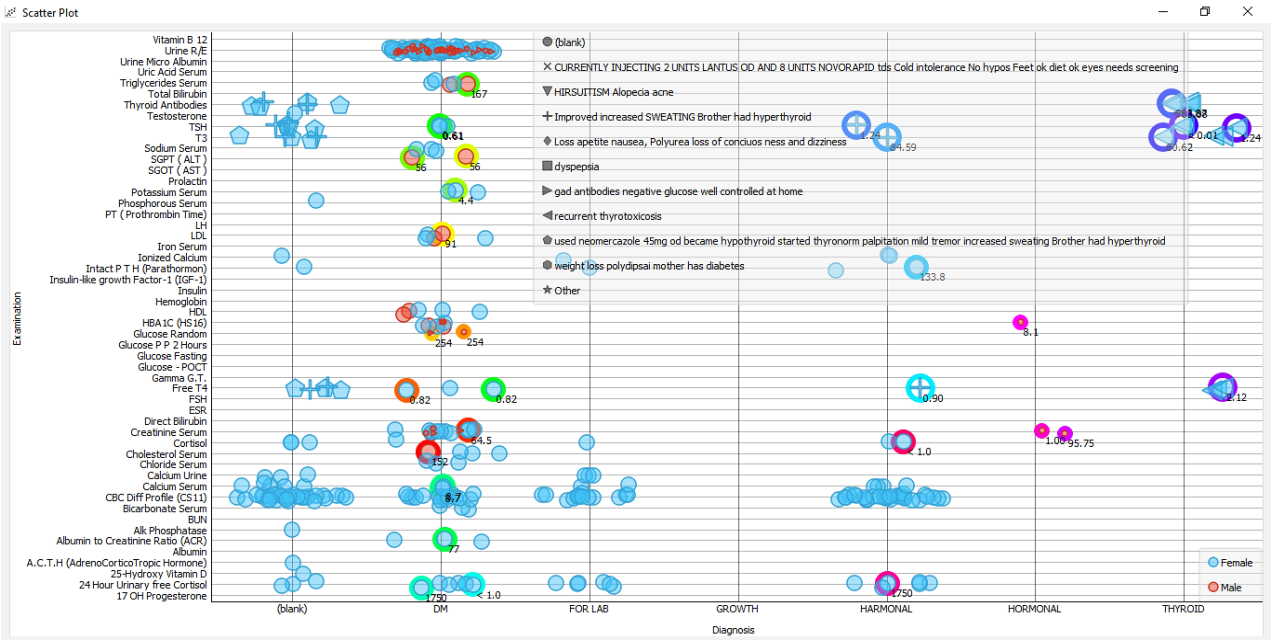

Fig. 18. Various test results for diagnosis of different forms of diabetes in males/females highlighted with Notes in Corpus100_DM_pts_2844.

kappa: 1.000 +/- 0.000 (micro average: 1.000)

	true Ster...	true Gest...	true DM	true COPD	true THY...	true SUB...	true Pred...	true Prim...	true Sec...	true Prim...	true HO...	true ANE...	true Park...
pred. Ste...	1098	0	0	0	0	0	0	0	0	0	0	0	0
pred. Ge...	0	352	0	0	0	0	0	0	0	0	0	0	0
pred. DM	0	0	18280	0	0	0	0	0	0	0	0	0	0
pred. CO...	0	0	0	1532	0	0	0	0	0	0	0	0	0
pred. TH...	0	0	0	0	3384	0	0	0	0	0	0	0	0
pred. SU...	0	0	0	0	0	450	0	0	0	0	0	0	0
pred. Pre...	0	0	0	0	0	0	22	0	0	0	0	0	0
pred. Pri...	0	0	0	0	0	0	0	10	0	0	0	0	0
pred. Se...	0	0	0	0	0	0	0	0	316	0	0	0	0
pred. Pri...	0	0	0	0	0	0	0	0	0	285	0	0	0
pred. HO...	0	0	0	0	0	0	0	0	0	0	1698	0	0
pred. AN...	0	0	0	0	0	0	0	0	0	0	0	78	0
pred. Par...	0	0	0	0	0	0	0	0	0	0	0	0	112
pred. Tox...	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. Co...	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. LU...	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 19. DMDL achieved 100% accuracy on Corpus14407_DM_pts_33185.

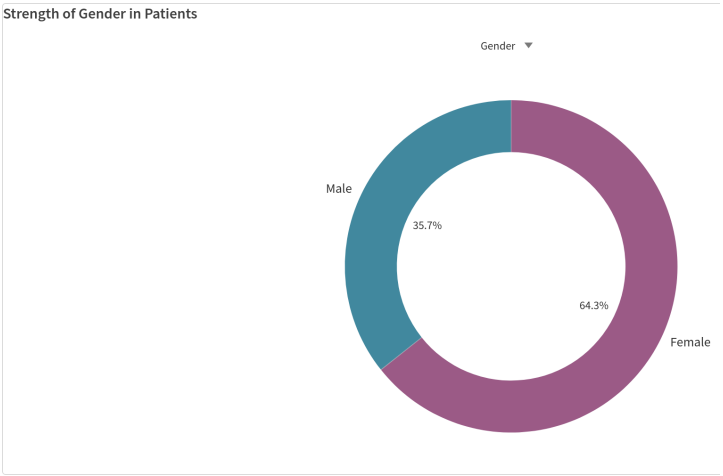

Fig. 20. Distribution of endocrine patients in Corpus14407_DM_pts_33185 with respect to gender.

It is seen again and again with respect to each corpus, the distribution of patients with respect to ages and mostly they lie in middle age bracket as in Fig. 21 here. In Fig. 22 again diagnosed class is seen more related with respect to age, lab results for recommended examinations, gender and allergy issues, respectively.

In the corpus with 33,185 records for 14,407 patients, Fig. 23 shows the distribution of all diagnoses in relation to the primary DM diagnoses that is exceeding in ratio.

Fig. 24 shows once again the complex corpus of 100 patients with 9304 records with maximum diagnoses where each diagnosis is shown in relation with average age of patients.

Fig. 25 shows the corpus of 14,407 patients with 33,185 records with diagnoses in relation to respective allergy where DM, Chronic Obstructive Pulmonary Disease (COPD), steroid induced diabetes, stroke and thyroid patients have Not Known Food Allergies (NKFA) and Not Known Drug Allergies (NKDA) mostly.

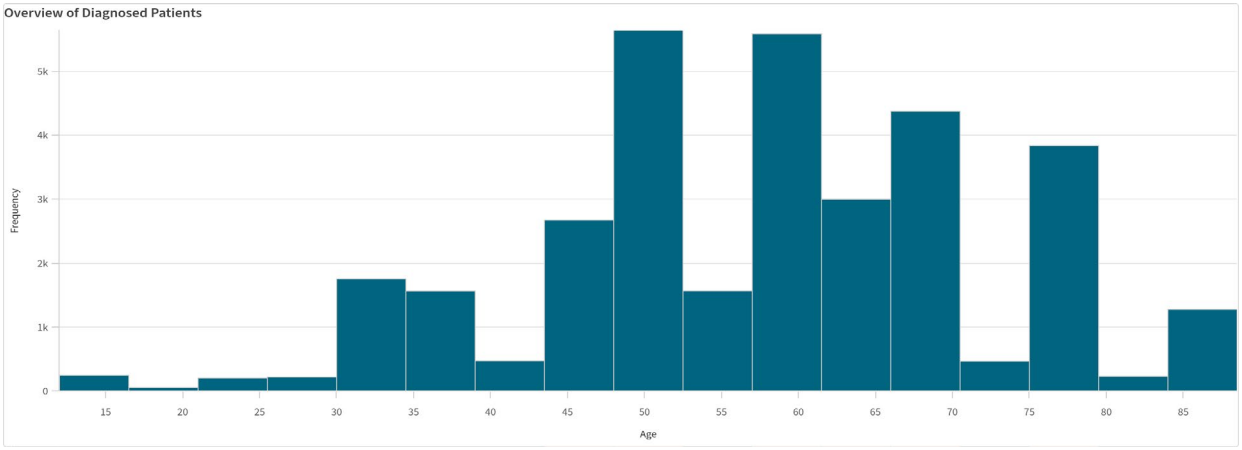

Fig. 21. Distribution of endocrine patients in Corpus14407_DM_pts_33185 with respect to age.

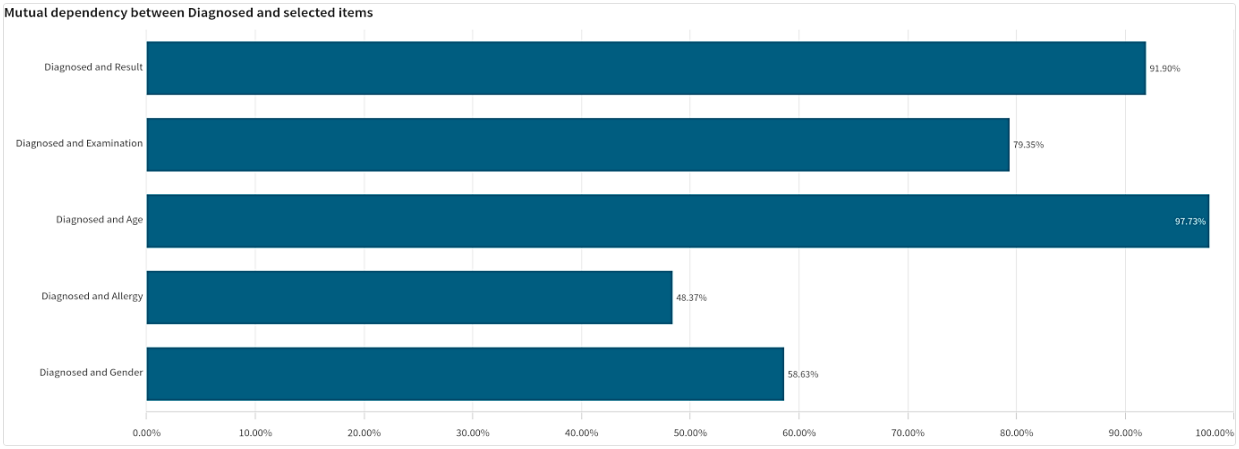

Fig. 22. Dependency of diagnosed disease on related feature sets (fields) in Corpus14407 DM pts 33185.

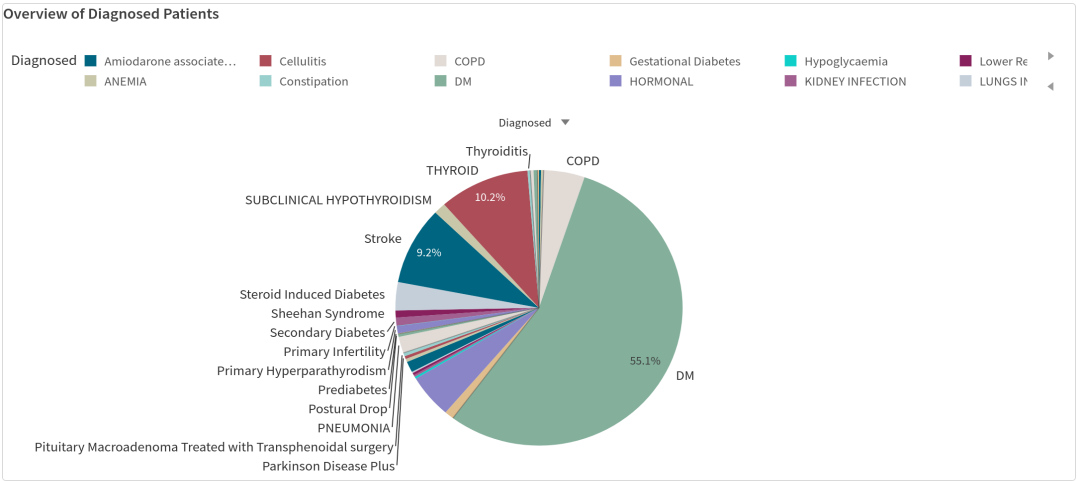

Fig. 23. Corpus14407_DM_pts_33185 having majority DM patient with other comorbidity diseases.

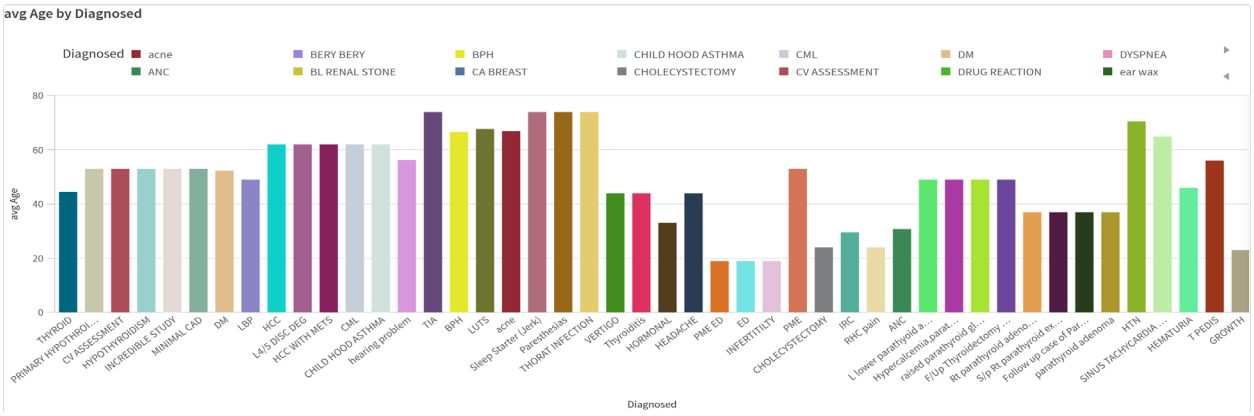

Fig. 24. Final Corpus100_DM_pts_9304 shows DM diagnosis with other comorbidity diseases that patients suffered.

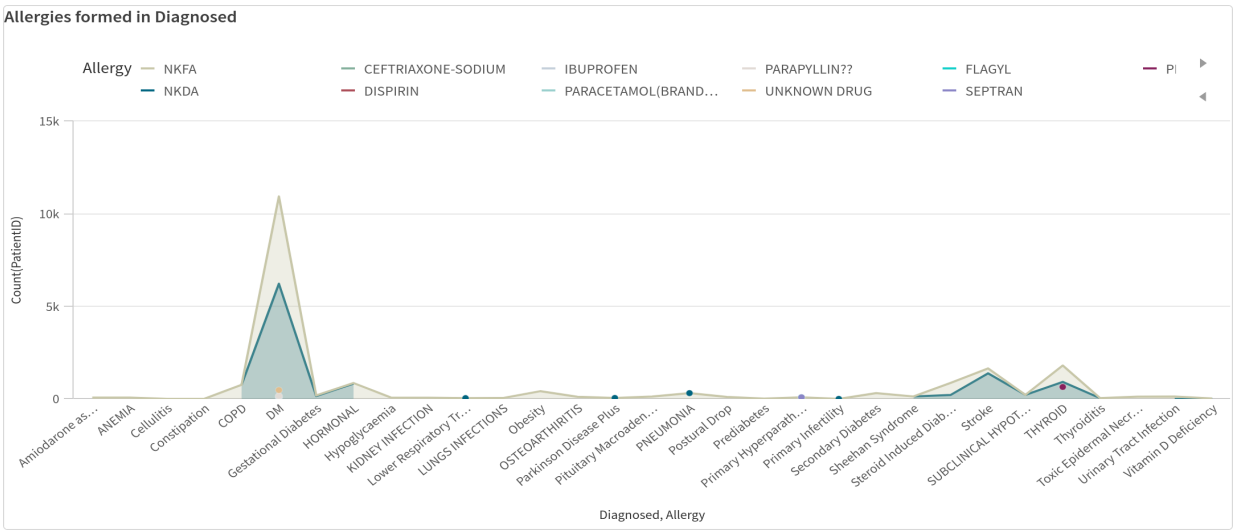

Fig. 25. Diagnosed endocrine patients in Corpus14407_DM_pts_33185 mostly relate to NKFA and NKDA.

Both proposed ensemble models on auto ML cloud platforms with deep learning heuristics; LMHFL and DMDL, were found significant for the diagnoses of DM and its comorbidities in generalized and interoperable environment. These models promise the accuracy, efficiency and speed for analysis when integrated with more high performance cloud platforms for big data in future.

V. Discussion

We used the proposed unified clinical data model to initiate work on unified clinical corpora design and development for collaborative research environment in SmartHealth. It forms the evaluation measure for healthcare analytics to keep accuracy level in check. In shared research environment we would be able to compare the strengths of proposed health analytics with the previous one [5]. Our first evaluation measure is to gather accurate and reliable health data for which we structured the proposed unified data model on ICD-10 and HL7 FHIR v4 format [32]. Secondly, the collection and development of standardized unified clinical corpora has started which would validate proposed analytical models for clinical use. Thirdly, the standardization makes these health analytics interoperable and generalized to be transferable over open-source cloud platform in compliance with HL7-FHIR standard [19].

The exhaustive experimentation and validation fine-tunes the analytics to give maximum accuracy on complex computation of the problem at hand (Table V). We evaluate its diagnostic accuracy based on the established embedding techniques [34, 37], biomarkers [38], probability, confidence interval, precision (usefulness of results) and recall (completeness of results) [6–16], or using some similar statistical tools for high performance computing [32–42].

RapidMiner Auto ML platform was even better to test and validate the custom model; DMDL. It had the distributive multi-label classification capability that showed results within seconds given our datasets. The diagnostic accuracy was 100% on multiple runs shown in form of confusion matrix.

Researchers would collaborate with clinicians and hospitals to judge the merits of the proposed algorithms and analytics to avoid conflict in its commercialization. Transparency and interoperability of algorithm design within controlled and secure environment may add cognition and engagement of patients and doctors.

There were four open-source cloud platforms that we experimented with to analyze datasets, design and build ML/DL heuristics for model/s development, test and

validate those. Table V focuses on main contributions of the proposed BDA models as depicted in the paper above. Here, an overview of achieved results is shown. QlikSense was good to statistically understand the data at hand, however, as the data grew in size we shifted to other tools; Google Colab, Orange and RapidMiner.

In Google Colab, DNN Bi-LSTM Sequential model was tested for diagnoses given multiple features. The accuracy increased as more features were introduced with maximum model accuracy of 0.9 and validation accuracy up to 1 given the nature of datasets (shown in Table V).

Orange gave us the platform to design, test and validate more high-performance model; LMHFL, that was even better when optimized with fast.ai deep learning library for textual analysis. It gave maximum accuracy of above 0.8 on multiple patients' diagnoses and for single patient with multiple diagnoses the accuracy was above 0.9.

TABLE V. JUSTIFICATION OF THE COMPLEXITIES FOUND IN VARIABLE SIZED CORPUSES HAVING VARIABLE NUMERIC AND TEXT FIELDS TAKEN FOR ANALYSIS GIVING DIFFERENT RESULTS ON PROPOSED ANALYTICAL MODELS/PLATFORMS

Open-Source Cloud platforms	Analytical Models	Corpus100_DM_pts_2844	Corpus100_DM_pts_9304	Corpus14407_DM_pts_33185
QlikSense	Statistical	QlikSense is good for understanding the data on statistical measures. On free platform it could only take up to 5000 records. Visualization results in; tree maps, scatter plots and histograms, were found better.	N/A	N/A
Google Colab	DNN Bi-LSTM Sequential Model	Model Accuracy = 0.4615 Validation Accuracy = 1 Diagnosed classes = 3 out of 5	Model Accuracy = 0.6 Validation Accuracy = 1 Diagnosed classes = 8 out of 65	Model Accuracy = 0.9 Max. Validation Accuracy = 0.8462 Diagnosed classes = 17 out of 32
Orange	LMHFL	Very good and explanatory patient specific visualizations were found with LMHFL that let us extract different associations in features for diagnosis of DM and its comorbidities. Max. accuracy for multiple patient profiles was; AUC of 0.982, CA of 0.921 and F1 score equaled 0.910.	On a larger dataset labels were too many to be correctly visualized for each patient. Correlations above 0.9 with maximum accuracy above 0.8, was achieved with Laplace. Rules were induced. AUC achieved was above 0.9 and F1 above 0.56. For single patient profile accuracy was above 0.9.	It would not take records above 30k due to limited resource capacity. Therefore, we split dataset into individual patient profiles having a single disease DM or its comorbidity as constant and Medicine, Test or Allergy as target variables. The model gave max accuracy above 0.9. Best confusion matrix, were seen with optimized fast-LMHFL for free text mining.
RapidMiner	DMDL	Deep learning auto model and our proposed DMDL both processed complete data with 100% accuracy.	DMDL showed 97.3% accuracy through confusion matrix. Total run time was only 13 seconds. The optimized DMDL with multi-label operator gave individual patient representations for DM and its comorbidities. Finally, it took processing speed ranging from 23s to 3mins to 6mins depending on the size and complexity of dataset provided and reached maximum of 100% accuracy in multiple runs.	Deep Multinomial Model with multi-label classification was applied that was fast. The model performed best with log loss of 0.08 tuned on different parameter settings. This balanced trained model was cross-validated with 10-folds using decision tree within that gave 100% recall and precision results with kappa equaling to 1 (Fig. 19).

VI. CONCLUSION AND FUTURE WORK

This research discussed and evaluated proposed analytics for semantic diagnostic intelligence keeping in view the complexity in view of endocrine EHR big data for DM and its comorbidity diseases. The proposed algorithmic models were; (i) DNN Bi-LSTM Sequential model, (v) LMHFL and optimized Fast-LMHFL, and (vi) DMDL optimized with multi-label operator.

The proposed analytics used deep learning heuristics integrated with NLP techniques that took proposed unified endocrine corpora for multiple diagnoses in a single patient with DM as primary disease.

The DNN Bi-LSTM Sequential model achieved maximum accuracy of 0.4615, 0.6 and 0.9 with maximum number of features selected with the increase in the size of corpus. The predicted diagnosed classes grew with size of corpuses where validation accuracies were 1, 1 and 0.846, 2, respectively. Accuracies signaled to the good quality of our corpora and the importance of free text clinical notes and practitioner comments from which the key attributes; condition, disease and medicine are extracted for accurate diagnoses.

LMHFL processed multiple patients' diagnostic profiles in Corpus100_DM_pts_2844 and Corpus100_DM_pts_9304. Maximum accuracy achieved was 0.7. Single patient profile was diagnosed with maximum accuracy of 0.9. Optimized Fast-LMHFL further increased the accuracy for single patient diagnoses.

Finally, DMDL optimized for multi-label classification processed all three corpuses in the proposed unified corpora with 100% diagnostic accuracy.

Data collection and curation was not at all easy. Firstly, finding MIS enabled Shifa International Hospital of repute

and getting its approval for quantitative as well as accurate qualitative data sharing was very challenging. Researchers had to make several visits to MIS to design and develop DFDs and ERDs in sync with HL7-FHIR standard. We observed the relevant doctors in their offices to understand the clinical process and data entry of patients while they discussed their history, symptoms lab test results with the doctors to come up with final diagnosis. Even after having the normalized datasets researchers put them through exhaustive data-warehousing and cleaning for de-normalized datasets that was fine enough to put through analytics. Still it was purified during analysis in parallel to reach its final form labeled with ICD-10-CM standard diagnostic codes that was DM_Comorbid_EHR_ICD10. However, these big EHR corpora are specifically for the diagnostics of endocrine patients suffering from DM and its comorbidities until it grows with time. Plus, as it is used to propose big data healthcare analytics it would not run on some traditional algorithms like; Support Vector Machine (SVM), etc. that are good to analyze small datasets.

The limitations are elaborated in Table VI to be overcome and refined in future.

Our proposed simulation design for hybrid SmartHealth cloud analytics framework may be used to integrate our analytics model as a relatively different approach compared to analytics techniques explored so far. It is being observed that several commercial companies like; IBM, Amazon, Microsoft, Google, Salesforce, etc. are competing to provide best AI healthcare services that are lacking yet. Our SmartHealth Cloud Analytics system would be able to leverage any such cloud platform capable of intelligent decision making on real-time in future. Our SmartHealth cloud analytics would be learning continuously from real-time patients' data coming in for checkups and examinations by tele-health clinicians for expert opinions. The tools and techniques would be upgraded for better results as mentioned in Table VI.

TABLE VI. RESEARCH LIMITATIONS THAT WOULD OPEN DOORS FOR THE RESEARCH IN FUTURE

Limitations	Future Work and its Benefits
We explored NLP domain for text classification and sequence embedding to tag diagnoses of DM and its comorbidities. But we know that our research is limited by time, scope and access to tools and techniques for exploration in NLP domain.	For NLP, we could only explore Bi-LSTM in keras tenserflow on Google Colab environment, fast.ai in Orange framework and text mining in RapidMiner used for auto ML but there are other algorithms like; BERT, Elmo, etc. that we would try in our future research.
Notes and PC text fields had lots of missing values that was a hindrance to diagnose.	In this limited time we took the opportunity to initiate the development of unified diagnostic corpora. The corpora would further grow to hold more diseases and would be purified. The issue of missing and null fields would also be taken care of with time and access to health data.
The manual annotation of the corpora, provided, is hence started using Spacy to come up with automated annotation.	Manual annotation takes time. We would further explore the area and may find better and faster ways of annotations.
Other domain that we touched in this study is visual analytics validated on Orange cloud framework. Before selecting Orange framework, we explored several other platforms like; Qlik Sense and Tableau, etc. But still there is room for exploration in multidimensional visualizations for clarity of results.	Python is known to be the best for deep learning in big data but for visualizations other tools like; "R", MATLAB, Power BI, etc. may help improve the results in future.

ETHICAL ISSUES

The authors have read and follow the data availability and ethical requirements for publication of this Data article and confirm that due to the sensitive nature of the real-time patients' EHR data analyzed in this study Patient is kept anonymous using PatientID and not patient name or MR. number. Hospital and other executive stakeholders are assured that it would only be shared for research purposes to the relevant community for its responsible and ethical use.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

SS: Data collection, methodology, investigation, writing and editing—original draft; QJ: Supervision, methodology, review original draft, project administration; MAS: Collaborating author, writing and review—original draft; HFA: Supervision, methodology, review—original draft, Resources. All authors had approved the final version.

FUNDING

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU250850].

REFERENCES

[1] L. Tamine and L. Goeuriot, "Semantic information retrieval on medical texts: Research challenges, survey and open issues," *ACM Comput. Surv.*, vol. 54, no. 7, pp. 1–38, 2021.

[2] Y. Si, J. Wang, H. Xu *et al*., "Enhancing clinical concept extraction with contextual embeddings," *J. Am. Med. Informatics Assoc*., vol. 26, no. 11, pp. 1297–1304, 2019.

[3] R. I. Dogan, R. Leaman, and Z. Lu, "NCBI disease corpus: A resource for disease name recognition and concept normalization," *J. Biomed. Inform*., vol. 47, pp. 1–10, 2014.

[4] S. Shafqat, Z. Anwar, Q. Javaid *et al*., "A unified deep learning diagnostic architecture for big data healthcare analytics," presented at the 2023 IEEE 15th International Symposium on Autonomous Decentralized System (ISADS), Mexico, IEEE, 1–8, 2023.

[5] S. Shafqat, S. Kishwer, R. U. Rasool *et al*., "Big data analytics enhanced healthcare systems: A review," *J. Supercomput*., vol. 76, pp. 1754–1799, 2020.
[6] I. Kavakiotis, O. Tsave, A. Salifoglou *et al*., "Machine learning and data mining methods in diabetes research," *Comput. Struct. Biotechnol. J*., vol. 15, pp. 104–116, 2017.
[7] F. Magrabi, E. Ammenwerth, J. B. McNair *et al*., "Artificial intelligence in clinical decision support: Challenges for evaluating ai and practical implications," *Yearbook of Medical Informatics*, vol. 28, no. 1, pp. 128–134, 2019.
[8] M. Afzal, M. Hussain, W. A. Khan *et al*., "Comprehensible knowledge model creation for cancer treatment decision making," *Comput. Biol. Med*., vol. 82, pp. 119–129, 2017.
[9] M. Krishnamurthy, P. Marcinek, K. M. Malik *et al*., "Representing social network patient data as evidence-based knowledge to support decision making in disease progression for comorbidities," *IEEE Access*, vol. 6, pp. 12951–12965, 2018.
[10] S. Neto and F. S. Ferraz, "Disease surveillance big data platform for large scale event processing," in *Proc. on the International Conf. on Internet Computing (ICOMP)*, 2016, pp. 89–95.
[11] N. M. S. Kumar, T. Eswari, P. Sampath *et al*., "Predictive methodology for diabetic data analysis in big data," *Procedia Comput. Sci*., vol. 50, pp. 203–208, 2015.
[12] L. Wang and C. A. Alexander, "Big data analytics in healthcare systems," *Int. J. Math. Eng. Manag. Sci*., vol. 4, no. 1, pp. 17–26, 2019.
[13] S. Shafqat, A. Abbasi, M. N. A. Khan *et al*., "Context aware smarthealth cloud platform for medical diagnostics using standardized data model for healthcare analytics," *Int. J. Adv. Comput. Sci. Appl (IJACSA)*., vol. 9, no. 7, 2018.
[14] J. Zhao, P. Papapetrou, L. Asker *et al*., "Learning from heterogeneous temporal data in electronic health records," *Journal of biomedical informatics*, vol. 65, pp. 105–119, 2017.
[15] M. Usama, J. Qadir, A. Raza *et al*., "Unsupervised machine learning for networking: Techniques, applications and research challenges," *IEEE Access*, vol. 7, pp. 65579–65615, 2019.
[16] A. P. Piotrowski, M. J. Napiorkowski, J. J. Napiorkowski *et al*., "Swarm intelligence and evolutionary algorithms: Performance versus speed," *Information Sciences*, vol. 384, pp. 34–85, 2017.
[17] S. Vaijayanthi and J. Arunnehru, "Deep neural network-based emotion recognition using facial landmark features and particle swarm optimization," *Journal for Control, Measurement, Electronics, Computing and Communications*, vol. 65, no. 3, pp. 1088–1099, 2024.
[18] S. Vaijayanthi and J. Arunnehru, "Dense SIFT-based facial expression recognition using machine learning techniques," in *Proc. the 6th International Conference on Advance Computing and Intelligent Engineering: ICACIE 2021*, Singapore: Springer Nature Singapore, 2022, pp. 301–310.
[19] S. Shafqat, A. Abbasi, T. Amjad *et al*., "Smarthealth simulation representing a hybrid architecture over cloud integrated with IoT: A modular approach," in *Proc. the 2018 Future of Information and Communication Conference (FICC)*, 2019, vol. 887, pp. 445–460.
[20] S. Kale, H. Tamakuwala, V. Vijayakumar *et al*., "Big data in healthcare: Challenges and promise," in *Proc. 6th International Conference on Big Data and Cloud Computing Challenges*, Kansas City, 2020, vol. 164, pp. 3–17. doi: 10.1007/978-981-32-9889-7_1
[21] S. Reddy, N. Ramya, K. Bharathi *et al*. (2021). Design and development of robust performance analysis of binary and multi-class models using azure machine learning. [Online]. Available: https://levantjournal.org/wp-content/uploads/2021/07/Design-and-Development-of-Robust-Performance-Analysis-of-Binary-and-Multi-Class-Models-Using-Azure-Machine-Learning.pdf
[22] J. Howard and S. Gugger, "Fastai: A layered API for deep learning," *Inf*., vol. 11, no. 2, 2020. doi: 10.3390/info11020108
[23] M. Raghu and E. Schmidt, "A survey of deep learning for scientific discovery," arXiv preprint, arXiv:2003.11755, 2020.
[24] H. Zhu, I. C. Paschalidis, and A. Tahmasebi, "Clinical concept extraction with contextual word embedding," arXiv preprint, arXiv:1810.10566, 2018.
[25] V. C. Kaggal, R. K. Elayavilli, S. Mehrabi *et al*., "Toward a learning health-care system–knowledge delivery at the point of care empowered by big data and NLP," *Biomed. Inform. Insights*, 2016. doi: 10.4137/BII.S37977
[26] T. Foley, L. Horwitz, and R. Zahran, "Realising the potential of learning health systems," *The Learning Healthcare Project*, vol. 101, 2021.
[27] D. D. Silva, F. Burstein, H. Jelinek *et al*., "Addressing the complexities of big data analytics in healthcare: The diabetes screening case," *Australas. J. Inf. Syst*., vol. 19, no. 2013, pp. S99–S115, 2015.
[28] N. Jankowski, "Approximation and classification in medicine with IncNet neural networks," presented at Machine Learning and Applications. Workshop on Machine Learning in Medical Applications, July 53–58, 1999.
[29] M. Grochowski and N. Jankowski, "Comparison of instance selection algorithms II. Results and comments," presented at International Conference on Artificial Intelligence and Soft Computing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.
[30] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Systems with Applications*, vol. 35, pp. 82–89, 2008.
[31] M. S. R. Nalluri, K. Kannan, and D. D. Roy, "Hybrid disease diagnosis using multiobjective optimization with evolutionary parameter optimization," *Journal of Healthcare Engineering*, 5907264, 2017.
[32] Administration-module-FHIR v4.0.0. 2018. [Online]. Available: https://www.hl7.org/fhir/administration-module.html
[33] S. Piri, D. Delen, T. Liu *et al*., "Development of a new metric to identify rare patterns in association analysis: The case of analyzing diabetes complications," *Expert Syst. Appl*., vol. 94, pp. 112–125, 2018.
[34] S. Shafqat, Z. Anwar, Q. Javaid *et al*., "NER sequence embedding of unified medical corpora to incorporate semantic intelligence in big data healthcare diagnostics," Research Square Preprint, 2024. doi: 10.21203/rs.3.rs-3148503/v2
[35] S. Shafqat, M. Fayyaz, H. A. Khattak *et al*., "Leveraging deep learning for designing healthcare analytics heuristic for diagnostics," *Neural Processing Letters*, pp. 53–79, 2021.
[36] S. Shafqat, Z. Anwar, R. U. Rasool *et al*., "Rules extraction, diagnoses and prognosis of diabetes and its comorbidities using deep learning analytics with semantics on big data," *Qeios*, 2023. doi:10.32388/67KZ7S.3
[37] L.Weber, M. Sanger, J. Munchmeyer *et al*., "HunFlair: An easy-to-use tool for state-of-the-art biomedical named entity recognition," *Bioinformatics*, vol. 37, no. 17, pp. 2792–2794, 2021.
[38] J. Li, B. Jiang, and J. P. Fine, "Multicategory reclassification statistics for assessing improvements in diagnostic accuracy," *Biostatistics*, vol. 14, no. 2, pp. 382–394, 2013.
[39] P. Nowicka, N. Santoro, H. Liu *et al*., "Utility of hemoglobin A 1c for diagnosing prediabetes and diabetes in obese children and adolescents," *Iabetes Care*, vol. 34, no. 6, pp. 1306–1311, 2011.
[40] A. Stranieri, A. Yatsko, H. Jelinek *et al*., "Data-analytically derived flexible HbA1c thresholds for type 2 diabetes mellitus diagnostic," *Artif. Intell. Res*., vol. 5, no. 1, 2015.
[41] X. Yang, J. Bian, Y. Gong *et al*., "MADEx: A system for detecting medications, adverse drug events, and their relations from clinical notes," *Drug Saf*., vol. 42, no. 1, pp. 123–133, 2019.
[42] D. Gannon, "A brief survey of cloud AI services," Technical Report, no. 12, 2018. doi: 10.13140/RG.2.2.26043.52004