A Unified Deep Learning Diagnostic Architecture for Big Data Healthcare Analytics

Sarah Shafqat, Zahid Anwar, Qaisar Javaid and Hafiz Farooq Ahmad

January 3, 2023

A unified deep learning diagnostic architecture for big data healthcare analytics

1st Sarah Shafqat
Department of Basic and Applied Sciences
International Islamic University (IIUI)
Islamabad, Pakistan
sarah.shafqat@gmail.com

2nd Zahid Anwar
Department of Computer Science and Sheila and Robert
Challey Institute for Global Innovation and Growth
North Dakota State University (NDSU)
Fargo. ND., USA
zahid.anwar@ndsu.edu

3rd Qaisar Javaid
Department of Basic and Applied Sciences
International Islamic University (IIUI)
Islamabad, Pakistan
qaisar@iiu.edu.pk

4th Hafiz Farooq Ahmad
Computer Science Department,
College of Computer Sciences and Information Technology (CCSIT)
King Faisal University
Al-Ahsa, 31982, Kingdom of Saudi Arabia
hfahmad@kfu.edu.sa

***Abstract*—Healthcare automation is evolving rapidly as can been seen in the recent popularity of e-health or digital health systems. The massive amount of health related data produced by these systems has given rise to the field of health informatics. The World Health Organization (WHO) decomposes SMARThealth as Standards-based, Machine-readable, Adaptive, Requirements-based, and Testable, and provides guidelines for digital health. Heterogeneous and big health data health that flows into the cloud requires considerations for uniformity of structure to allow for interoperability and generalizability for universal use and analysis. This research proposes a deep-learning architecture for disease diagnosis that considers Diabetes Mellitus (DM) as a case study. Three corpuses containing DM patient data are considered which are prepared and processed using extensive data warehousing techniques and labeled with ICD-10-CM diagnostic codes. Extraction of desired health data is through a unified data model for healthcare that is in compliance with HL7 FHIR v4.0 schema. Our contributions are two-fold: First, three big data cloud analytical models are proposed and validated on the unified corpora and second the maximum possible diseases specific to a single or multiple DM patients have been diagnosed with a 100% accuracy using deep multinomial/multi-label distribution learning (DMDL).**

***Index Terms*—deep learning, diagnostic architecture, healthcare analytics, HPC, big data**

I. Introduction

The World Health Organization (WHO) promotes quality health services for the general public and encourages the use of digital health systems [1]. The conception of health informatics [2] was made in the 1970s during World War II in which the role of computers for medical diagnoses was recognized as a viable option. Statistical data analysis with logic and probabilistic reasoning indicates that medical decision making is possible through computation. Steps taken to digitize health for the well being of the population at large still requires participation of all stakeholders and captures the confidence of hospitals, doctors and patients. Lately, this research has gained impetus with the availability of health care big data and technologies such as natural language processing (NLP). Further it has materialized into learning healthcare systems (LHS) pioneered by Mayo Clinic, USA and provisioned by WHO [3]. This research tackles several related research problems centered around document handling for knowledge processing. The clinical processes in LHS that have so far been catered to are hyperlipidemia, atrial fibrillation and congestive heart failure (CHF) out of a total of 115 that were initially conceptualized. Moreover LHS are being designed and adopted at several regions but these efforts require convergence [4]. This research is a first step towards addressing the need for a universal health platform that benefits the general public. The concept of a SmartHealth cloud was proposed in 2018 [5], [6], to bridge many concurrent efforts together and make available healthcare facilities to patients closer to their locations making the process more patient-centric. SmartHealth is gaining momentum with the introduction of digital devices in our daily lives that has connected the world's population [7]. WHO realized the breakthrough and gave SMART guidelines for transformation of digital health with a five-step pathway that is; **s**tandards-based, **m**achine-readable, **a**daptive, **r**equirements-based and **t**estable [1].

Current widely used commercial telemedicine systems are missing key analytical features burdening healthcare practitioners in having to spend more hours managing healthcare data online and patients not knowing how to find a matching doctor among other issues. This research therefore proposes big data analytics for diagnoses that is trained on real-time electronic health records (EHR) data. These big data health analytics would be compatible with SmartHealth cloud platforms for interoperability and may be used by the general public.

The real-time EHR big data fetched was of endocrine patients. It is widely realized that within the current lifestyle adopted worldwide, increasing isolation and unhealthy diet is making populations susceptible to Diabetes Mellitus (DM). This disease becomes chronic and often leads to patients being diagnosed with several other relatable diseases when left undiagnosed or mismanaged. This research aims for DM's timely diagnosis. Therefore the case study involved acquiring diagnostic big EHR datasets of DM patients that were also suffering from other comorbidity diseases. Most hospitals in Pakistan are not well equipped to hold EHR data. That is why a challenge faced was getting big EHR data of endocrine patients in parts from 100 to 14407 DM patients. This dataset not only had diagnostics data for DM but also included diagnoses of around 100 comorbidity diseases that co-occur in a DM patient with time.

Performing big data SmartHealth cloud analytics is challenging in that there exist limited, if any, unified standard diagnostic frameworks. A standard interoperable diagnostic architecture would house the analytics that were trained on dataset that followed health level seven (HL7) fast healthcare interoperability resources (FHIR) v4.0 schema [8]. Therefore, the next challenge was the standardization and unification of the big EHR data to make it interoperable for SmartHealth cloud for generalized diagnostics. This involved following the HL7 FHIR v4.0 standard schema and labeling the diagnoses with ICD-10-CM codes for universal use. Our proposed unified data model mapped on the HL7 FHIR v4.0 standard schema served the purpose of extracting EHR normalized datasets in excel sheets. These datasets were transformed using extensive data warehousing techniques into three flat tabular excel sheets of variable sizes and features. The corpuses/datasets labeled with ICD-10-CM diagnostic codes were composed in corpora to become the input to validate the proposed big data analytics as illustrated in Figure 1.

Deep Learning heuristics is efficient at analyzing big data but is challenged in understanding unstructured free text data in any given context such as in healthcare. To address this over fifteen phenotypical diagnostic features were explored to select the best feature sets for diagnostic rule mining. In these selected diagnostic phenotypes or features if the data field had free text such as clinical notes or practitioner comment and observations then appropriate NLP techniques were applied for text mining. These features become the nodes that initiate any analytical model based on advanced Machine Learning or Deep Neural Networks (DNN). The analytical model that is to be trained on the corpora with huge number of records in parts from multiple visits of 100 to 14407 DM patients diagnosed with other co-existing diseases would need high performance computing (HPC) platform. We chose to train our proposed **DNN Bi-LSTM sequential model** on Google Colab first but found that it could only diagnose a few diseases based on the size of data and the selected feature sets with 90% maximum accuracy. The second analytical model that we proposed was **Louvain Mani-Hierarchical Fold Learning (LMHFL)** and its optimized version **Fast-LMHFL** that integrated fast.ai deep learning library for text mining on the Orange framework to give best visualizations. The resultant inferences showed associations between features and multiple diagnoses for single as well as multiple DM patients suffering from other comorbidity diseases. Fast-LMHFL was found efficient at comprehending free text fields of clinical notes and practitioner comments. But, we found it limited to take corpus having records above 10k. Finally, we custom designed and proposed the third model on RapidMiner auto ML framework; **Deep Multinomial/Multi-label Distribution Learning (DMDL)** that processed all three corpuses in seconds and achieved 100% diagnostic accuracy shown in confusion matrix.

The unified data model that we propose will help researchers to extract the health data in their specific regions for any clinical decision making problem. This would grow the unified corpora that we initiated with the focus on endocrine patients; 'DM_Comorbid_EHR_ICD10'. Our proposed high-level diagnostic framework and big data healthcare analytical models are adaptive and reusable for future research.

The rest of the paper encompasses section II which is allocated to discussing previous research for similar systems that became the foundation to carry forward our research objective. Section III defines the limitations and strengths in the previous works. In section IV, we provide a description of the proposed solutions and section V gives an overview of results gathered from experiments. Finally, in section VI, we conclude our findings and pave the way to strengthen our research in future.

II. Previous Research Studies

An in-depth study of big data healthcare was conducted in [5]. This study is useful in providing the standard conventions and organizational bodies working towards integrating healthcare with technology. Some researchers have used deep neural networks (DNNs) to enhance the capability of machine learning algorithms for processing images like; Convolutional neural network (CNN), Bayesian networks [9], MLP [10], K-Means [11], Artificial Neural Networks (ANN) and decision tree by enabling the processing of vectored and complex datasets that initially worked on scalar datasets only. Transformation of traditional clinical practices into digital form causes a doubt of uncertainty on the accuracy of the prediction. Generalized performance is said to be achieved when model's performance on training and test set has minimal gap. It is achieved by training models on patient profiles that have similar symptoms, lab tests and diagnosis and then checking when the desired prediction performance has been achieved [12], [13].

Lately, the application of DNNs [14], [15] is seen in the healthcare domain to diagnose patients with diabetic retinopathy (DR) using the scans found in medical EHRs. The results and analysis show the diagnostic precision matching to that of qualified clinicians. Still it is important that the final clinical decision and recommendation must consider possible uncertainties inherent in machine enabled diagnosis that may put patients' health at risk. Bayesian DNNs are found computationally expensive and trade performance to reduce the uncertainties in diagnosis using dropout, regularizers or

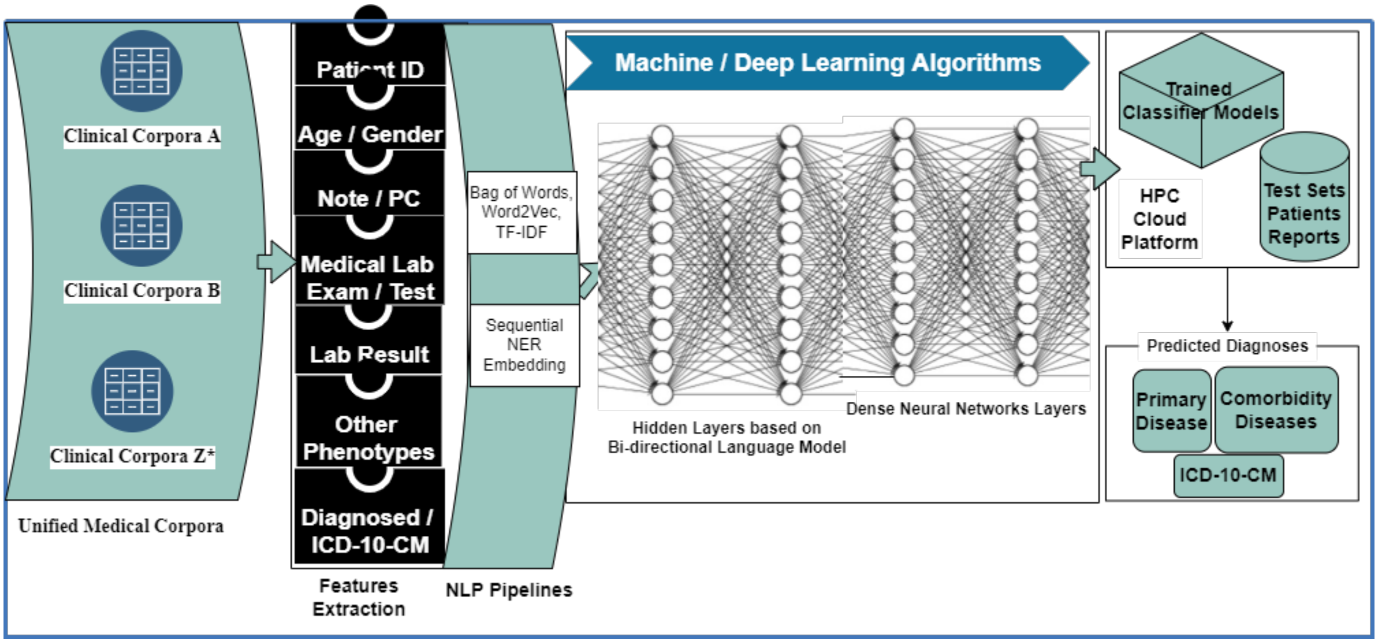

Figure 1. High-Level Deep Learning Diagnostic Framework for cloud.

batch normalization. The alternative to Bayesian is ensemble DNNs where each node is initialized at random to sample diverse accurate predictions improving single network performance but at the cost of increased training and interpretation. Sathiya et al. [15] evaluated mainly two classes of DR diagnosis; mild and moderate while Ayhan et al. [14] predicted uncertainty distributions on five levels of DR; (i) No DR, (ii) Mild DR, (iii) Moderate DR, (iv) Severe DR and (v) Proliferative DR. Ayhan et al. [14] emphasize that even experienced doctors find risk of uncertainties in diagnosis and therefore digitization is found promising to reach the final decision. They conducted experiments using test-time data augmentation (TTAUG) in DNNs to be intuitive through a data-driven approach. The diagnostic uncertainty was validated by matching the proportionality of clinicians' disagreements to the rate of uncertainty predicted in diagnosis. Two datasets were considered; the first one containing retina images from a Kaggle Competition and second one called the Indian Diabetic Retinopathy Image Dataset (IDRiD) for analysis on a severity scale set by International Clinical Diabetic Retinopathy. A modified version of CNNs were employed which introduce an activation function called softmax used in the last layer. Other fully connected layers use parametric rectifier logical units (PReLUs) and an additional layer. This CNN was applied on the DR image datasets modified to fully connect all layers with an additional layer before softmax. The algorithm used PReLUs as an activation function on first fully connected layers in the stack and the additional layer. The maximum and average pooled features were extracted using batch renormalization (BReN). The output of 512 features was then fed into the softmax 5-way fully connected layer for classification. Cross entropy loss was used to train for 500000 iterations. Stratified sampling was applied with increasing batch size after 50000 iterations starting with 20 images in a mini-batch. An L2 regularizer was used in the first stack of layers that adopted L2 regularization measure in the fully connected layer. The network adjusted itself to the distribution of classes. Data augmentation was performed at every iterated cycle but was not found enough for predicting uncertainty estimation and TTAUG was proposed for deterministic classes. A modified residual neural network (ResNets) was used to act as DNNs and results were gathered for evaluation using TTAUG and validation was performed through experts' feedback. Ensemble predictions were found to be robust even if they lacked in evaluating the uncertainties in DR diagnosis. The model was tested with TTAUG to improve the generalization gap between training and test data and the cost was redeemed for higher T on getting lowered discrimination performance. The underconfident results yielded through ensemble were found to be highly accurate.

In the year 2000, Johnson et al. [16] recognized the need for a unified real-time data mining framework that would ensure the quality and completeness of data taken as input to validate analytics. Research was being carried out in solitude and no uniform data format was available to enable uniformity in healthcare analytics. Till date the robustness of data handling and analysis with uniformity is still lacking.

Several open-source services and infrastructures are present in the form of MapReduce and Apache Spark implementation on Hadoop Distributed File System (HDFS), RapidMiner, Orange, etc. that house several ML algorithms known to speed up computing and processing of big complex datasets [17]. Still several challenges are linked to healthcare big data modeling due to high dimensionality and complex business processes and flow of information. Advancements have been made to standardize the medical vocabulary in form of systemized nomenclature of medicine clinical terms (SNOMED-CT), medical subject headings (MeSH), nuclear magnetic resonance spectroscopy for metabolomics data markup language (nmrML) and the international classification of diseases (ICD). These standard medical conventions help to annotate and form ontologies related to medical context [17].

All these tools and techniques can be considered as foundational building blocking for achieving a mature learning healthcare system (LHS) as the one that is under development at Mayo clinic, USA, since 2001 [3]. Several NLP techniques have been employed since then for high-level information extraction through developing of dictionaries that define the patterns or terms to be extracted, normalization of text that is mapped to the target concept and regular expressions that are used to set the rules between the first two components. NLP enriched unified data platform (UDP) is still limited in its capacity to empower real-time big data analytics to the point of care at Mayo Clinic. Recently, Mayo Clinic launched a clinical support system known as Mayo Expert Advisor (MEA) with a front end web application named AskMayoExpert (AME) which appears quite manual at the time. The challenges involve the rapid processing of unstructured clinical notes and other resources for personalized recommendations to patients [3]. Motivated by the potential of LHS, similar projects are being taken in other facilities at regional and organizational levels [4].

III. Limitations and Strengths of Previous Systems

This research builds upon previous research works starting from the realization of healthcare informatics [5], [6] to the beginning of the concept of SmartHealth [5]. A SmartHealth cloud needs to be developed for the unification of various healthcare services that can deliver point-of-care facilities to patients globally [7].

Deep learning [9]–[11], [14] is allowing for a speed up in processing where there is large or complex data and majorly in the classification and analysis of images as in retinopathy diagnosis. In order to perform meaningful analysis of clinical data NLP needs to be integrated for text mining and understanding of context. Further, the hybridization or ensemble of multiple ML algorithmic models strengthens the analytical model as in [18].

Till now DNNs are mostly being applied on healthcare images or fuzzy datasets and need to incorporate text mining capabilities for better inferences.

The challenge remains in the unification and standardization of all the efforts for the digitization of health care services [17].

IV. Proposed Solutions

In early 2010, Harvard Medical School and Boston's children hospital started an initiative to design a health project for interoperability that could be run across other medical applications/systems. This project in 2013 adopted the HL7-FHIR openly available draft standard for interoperability and was renamed as Substitutable Medical Applications and Reusable Technologies on FHIR (SMART on FHIR) [19]. Several prototypes were demonstrated in a conference in 2014 to establish its feasibility for commercial use. Later health projects include the LHS [3], clinical decision support systems [20] and analytics [17] but were mostly built in silos leading us to propose the concept of a SmartHealth context-aware hybrid cloud platform that integrates big data analytics [5]–[7], [20]. The platform ingests data from a variety of sources such as EHRs, social media, websites, documents and internet of everything attached to monitor patient's health through wearable devices and sensors. Recently, WHO has given a new definition to SMART as Standards-based, Machine-readable, Adaptive, Requirements-based, and Testable, to give guidelines for digital health [1].

A detailed analysis of all these solutions lead us to initiate this research on a standard interoperable diagnostic architecture that could host SmartHealth cloud analytics focused on automated disease diagnostics. Such an architecture requires the following artifacts:

- A high-level deep learning diagnostics architecture for the cloud
- A unified clinical data model
- A unified medical corpora
- Big data healthcare analytical models

A. High-Level Deep Learning Diagnostic Architecture

The proposed architecture of the proposed system is illustrated in Figure 1. The architecture has been designed to be adaptable so as to fulfill all clinical purposes over a cloud infrastructure. The five artifacts needed to train big data healthcare cloud analytical model(s) for clinical decision making are reflected in the model. The incoming heterogeneous data adapts to the proposed deep learning architecture for analytics through the unified data model that is in compliance with HL7. This transformed data or **corpora** is inputted from the left for **features extraction**. The right features set is passed through an **NLP pipeline** to tag or annotate any free text present. The extracted feature inputs then inserted into a language model for semantic understanding and iterated within a **DNN model** over an HPC cloud platform. The resultant **classifier model** is trained to predict the primary diagnosis with other secondary or comorbidity diagnoses that co-occurred in a patient with time.

B. Unified Clinical Data Model

A unified clinical data model that is in compliance with HL7 FHIR standard schema is required to shape the input data for

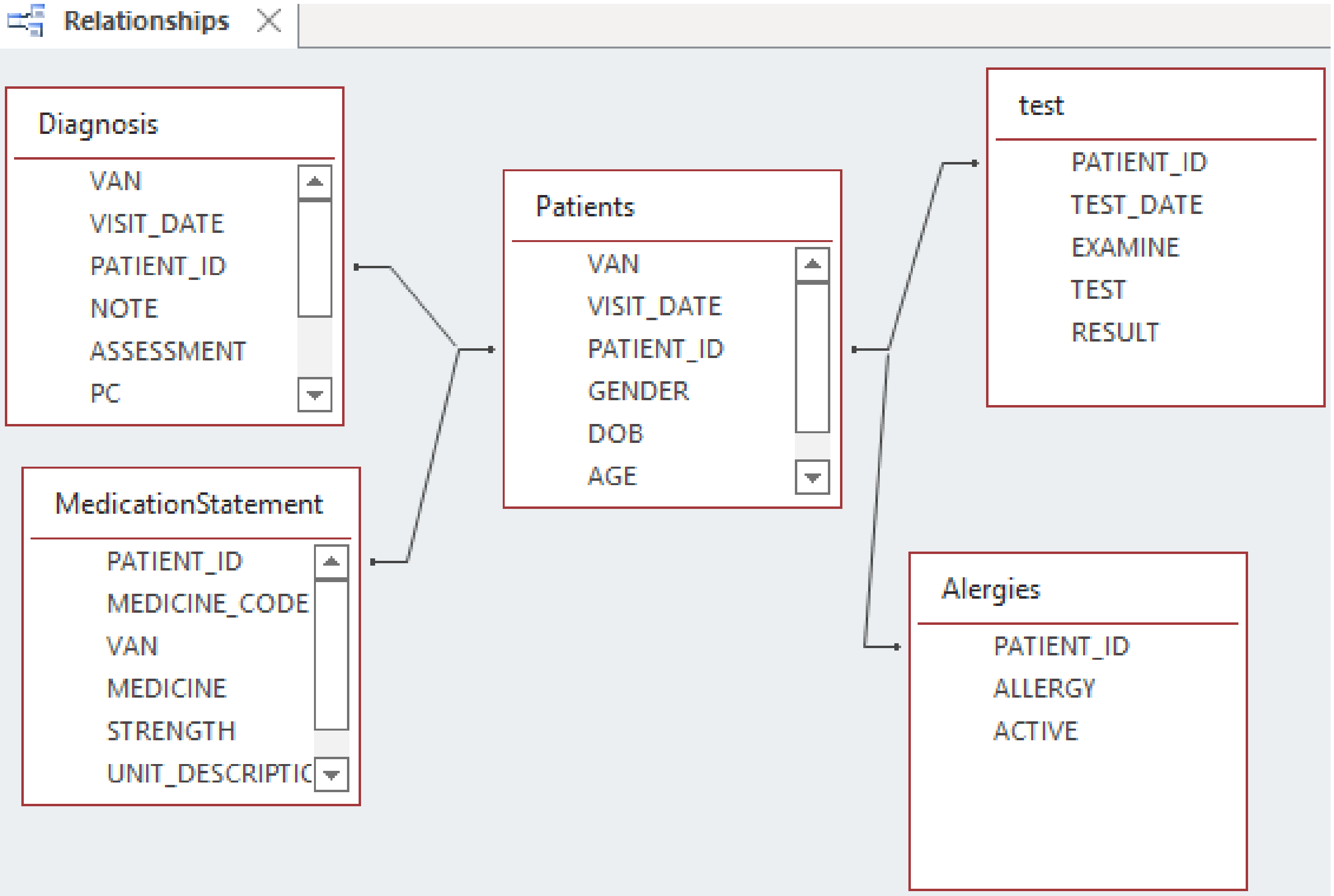

Figure 2. A unified diagnostic model depicted as an entity-relationship diagram (ERD) in compliance with the HL7 FHIR v4.0 standard schema for providing interoperability over the cloud.

interoperability over the cloud. The model consists of entities that allow for an accurate diagnoses and is designed and built in MS Access in form of an entity relationship model (ERD) depicted in Figure 2. The model may be expanded as per clinical needs and the usecase at hand.

C. Unified Medical Corpora

Data is fetched from various health sources and transformed into multiple corpora/corpuses and referred to by a naming convention. For an initial naming these are referred to as clinical corpora A, B and so on in Figure 1.

The preferred resource in our case was EHRs maintained in hospitals. The desired datasets were extracted against the entity tables drawn from the data model given above. The data collected was diagnostic information of endocrine patients having diabetes mellitus (DM) as primary disease and comorbidities as secondary diseases formed through complications or mismanagement over time. Datasets of 100 to 14407 DM patients was provided in parts by Shifa International Hospital, Pakistan. These normalized datasets were de-normalized to form three flat tabular datasheets after going through rigorous data warehousing techniques for pre-processing and cleaning. The three finalized datasets or corpuses became one unified corpora; **'DM_Comorbid_EHR_ICD10'**, following a naming convention showing that the corpora extracted from EHR is of DM patients with comorbidity diseases labeled with ICD-10 codes (Table I). Corpuses are also named with respect to the number of patients and the number of records or instances they contain.

D. Big Data Healthcare Analytical Models

The performance and speed of big data healthcare analytics is seen to improve with the introduction of deep learning heuristics [18]. Auto ML capability in RapidMiner was used to compare the performance of different ML algorithms on the corpora and deep learning was found to be best among others.

The findings helped propose three analytical models using deep learning heuristics within open-source cloud platforms; Google Colab, Orange and RapidMiner Auto ML frameworks.

a) DNN Bi-LSTM Sequential Model: We extended the features list to test and validate the model for its adaptability to varying number of features and records. The model was trained on all three corpuses in the corpora. The performance is evaluated based on the three feature sets and variable sizes. The best accuracy in diagnostic results was seen where there

Table I
DENORMALIZED META DATA SETS IN RAW FORM TRANSFORMED INTO CORPORA; DM_COMORBID_EHR_ICD10 CONTAINING THREE CORPUSES

Datasets	Dataset 1: 3650 instances of 100 patients	Dataset 2: 15696 instances of 100 patients	Dataset 3: 87803 instances of 14407 patients
Corpora	Corpus100_DM_pts_2844	Corpus100_DM_pts_9304	Corpus14407_DM_pts_33185
Features	PatientID, Age, Gender, VAN, Appointments, Note, Test Date, Examination, Test, Result, Assessment, PC, Diagnosed, ICD-10-CM	PatientID, Gender, Age, VAN, Appointments, Note, Test_Date, Examination, Test, Result, Assessment, PC, Diagnosed, ICD-10-CM	PatientID, VAN, Visit_Date, Age, Gender, Examine, Test, Result, Allergy, Note, PC, Medicine, Strength, Unit_Description, Days, Diagnosed, ICD-10-CM

were more features and the size of corpus was large and not complex.

b) Louvain Mani-Hierarchical Fold Learning (LMHFL): We validated the performance of this model on the unified corpuses that were formed on HL7 FHIR v4.0 schema for generalizability and interoperability. It processed the corpus; **Corpus100_DM_pts_2844**, containing 100 DM patients having around 2800 instances from multiple visits and majorly diagnosed with thyroid and hormonal diseases. The best visualizations are seen through multidimensional scaling (MDS), scatter and hierarchical graphs. The association rules were induced for important diagnostic features and classification accuracies (CA), area under graph (AUC), f1-score, precision, recall and confusion matrices (CM) were calculated. The other corpus; **Corpus100_DM_pts_9304**, that had records from the same 100 DM patients were larger in size due to the exhaustive list of around 65 comorbidity diseases that affected these patients. This corpus had over 9000 instances because a single patient had multiple visits and having multiple diagnoses records of DM and other co-existing comorbidity diseases. LMHFL was able to process the corpus and enabled the drawing of interesting inferences however the accuracy was not as good. We also analyzed single patient records that showed close associations between DM and comorbidity diseases with associated features. The third corpus; **Corpus14407_DM_pts_33185**, consisting of data from 14407 DM patients had more than 30000 instances which led to failure in processing the data in its entirety. Since this corpus contained records for a single diagnosis per DM patient we had to analyze a single patient that had constant diagnostic class with other target classes like; medicine in use, recommended laboratory tests or examinations and allergies formed against symptoms and the given diagnosis.

c) Fast-LMHFL: LMHFL when integrated with the fast.ai deep learning library [21] allowed for mining of free text found in the clinical notes and practitioner comments present in the corpuses. This improved the diagnostic results that was transparent through confusion matrices. For instance we observed that in LMHFL, a female DM patient who suffered from chronic breast cancer and other comorbidity diseases was only predicted for breast cancer. LMHFL failed to separately diagnose her for DM and other diseases she was suffering from while having breast cancer. When the same patient profile was validated with Fast-LMHFL it was easily differented between the diseases.

d) Deep Multinomial/Multi-label Distribution Learning (DMDL): DMDL was custom designed in the RapidMiner auto ML framework to mitigate limitations in the basic auto model that hindered the accuracy and processing speed. Patients were processed in batches and target roles were set for 16 features where 'Diagnosed' was set as target label, was predicted and then clustered. 'ICD-10' was also set as target label class. tanhdropout and softmax were used as activation functions and L1/L2 as regularizers to avoid overfitting. The model was cross validated with 10-folds using the decision tree algorithm and depicted a 100% accuracy. We ensembled it with a multi-label operator to further optimize it for reduced log loss error. In the optimized version all 16 features were set as target labels with 'Patient-ID' as batch parameter. It successfully processed the entire corpora with a 100% diagnostic accuracy enabled with multi-label classification operator.

V. RESULTS GATHERED

This section details the results the accuracy and performance of the diagnostics using the deep learning heuristics integrated with NLP and trained on the proposed unified corpora.

The DNN Bi-LSTM Sequential model built on DNN tensorflow.keras sequential model ensembled with bidirectional LSTM language model was first tested on **Corpus14407_DM_pts_33185**; corpus of 14,407 DM patients having above 30,000 records from multiple visits [22]. These DM patients with comorbidity diseases have a free text field called 'Note' and the target 'Diagnosed' class to make a albeit sequence (x,y) for named entity recognition (NER) tagging of right disease affecting the patient. The 'Note' freetext field is sparse in nature and results in zero model accuracy. This failure led us to make other possible albeit sequences ('Test', 'Diagnosed') or ('Exam', 'Diagnosed') that gave a maximum model accuracy of 0.56 and model validation accuracy of 0.88 [22]. But, these results were limited and were only based on one feature; 'Test' or 'Exam', and was fit to diagnose only DM out of a total of 30 diseases in the corpus. The model was further refined by validating it on all three corpuses taking multiple features. The accuracies indicated a good corpora quality and emphasize the importance of analyzing free text clinical notes and practitioner comments from which the key attributes; condition, disease and medicine are extracted. It finally achieved a maximum accuracy of 0.46, 0.6 and 0.9 with a maximum number of features selected with the increase in the size of corpus. The predicted diagnosed classes grew with size of corpuses where the validation accuracies were 1, 1 and 0.85 respectively.

Table II
DIAGNOSTIC ACCURACIES ON PROPOSED ANALYTICAL MODELS/PLATFORMS TRAINED ON VARIABLE SIZED CORPUSES

Open-Source Cloud platforms	***Analytical Models***	***Corpus100_DM_pts_2844***	***Corpus100_DM_pts_9304***	***Corpus14407_DM_pts_33185***
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	Highly explanatory patient specific visualizations. LMHFL allowed extraction of different associations in features for diagnosis. A maximum accuracy for multiple patient profiles was; AUC of 0.98, CA of 0.92 and F1 score of 0.91.	On a larger dataset labels were too many to be correctly visualized for each patient. Still correlations were found with maximum accuracy of 0.7 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.	The data was too big for the model to process due to limited resource capacity. The dataset was therefore split 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 a maximum accuracy above 0.9.
RapidMiner	DMDL	Deep learning auto model and our proposed DMDL both processed the entire data with 100% accuracy.	Achieved 97.3% accuracy through confusion matrix. Optimized DMDL with multi-label operator gave individual patient diagnosis. Processing speed ranged 23s to 6 mins depending on dataset size and complexity a maximum of 100% accuracy was seen in multiple runs.	The model was fast with the best performance seen with a log loss of 0.08 tuned on different parameter settings. This balanced trained model was cross-validated with 10-folds using a decision tree within that gave 100% recall and precision results with kappa equaling 1

LMHFL a form of DNNs was designed as graph neural networks (GNNs) and validated on the Orange framework for explainable visualizations. It was found very useful for drawing inferences for DM diagnostics and its associations with other comorbidity diseases [18]. We also tested it on other datasets for diagnoses of various types of dengue fever and covid-19 patients that reflected its efficiency to analyze big data sets. LMHFL processed multiple patients' diagnostic profiles in Corpus100_DM_pts_2844 and Corpus100_DM_pts_9304. The maximum accuracy achieved was 0.7. Single patient profile was diagnosed with a 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.

The detailed comparison of the diagnostic results is shown in Table II.

VI. CONCLUSION AND FUTURE WORK

Learning healthcare system (LHS) in 2016 was provisioned as a project of WHO that comes under the umbrella of United Nations SDG3 agenda. The work is ongoing in mayo clinic in USA and similar projects have taken pace in other facilities as well.

This learning mechanism is further extended by the concept of context aware SmartHealth to facilitate patients at their location using the cloud as a platform. This infrastructure would be empowered by big data healthcare analytics that involves advanced AI and NLP techniques ingrained to analyze several heterogeneous types of medical data that comes from EHR systems built in hospitals, social media and wearable devices. These analytics would help clinicians to predict, diagnose, treat and manage the patient.

Uniformity is required to structure the health data coming in from different sources over the cloud. The input data therefore has to be transformed to comply with HL7 FHIR standard schema for interoperability and generalizability. Clinical mechanisms such as diagnostics are considered in this paper which follow ICD-10 codes for universal use. Unified medical corpora developed in this work will grow to contain clinical data from different parts of the world with time.

Different Big Data Healthcare Analytics are proposed on top of high-level deep learning architecture that would be platform independent. The analytical models trained in Colab, Orange and RapidMiner would be able to integrate with other HPC cloud platforms such as AWS. In the future these analytics integrated over HPC cloud platforms would fasten the speed of processing to enable continuous data intake for distributed computing.

We validated our proposed analytics on records ranging from 2500 to 10,000 and over 30,000. Deep learning analytics processed 10,000 records with some selected features with considerable success. As the records grow over 30,000 with increased diagnostic phenotypes/features Deep Multinomial/Multi-label Distribution Learning (DMDL) was efficient and showed 100% accuracy. Therefore, it is clear that to process continuous streaming heterogeneous healthcare big data analytics will need to enable distributed deep learning heuristics that would need multiple GPUs to hold data in chunks or batches for fast processing.

ACKNOWLEDGMENT

The authors highly regard the sharing of EHR endocrine diagnostic data by Shifa International Hospital, Pakistan.

REFERENCES

[1] G. Mehl et al., "WHO SMART guidelines: optimising country-level use of guideline recommendations in the digital age," Lancet Digit.

Heal., vol. 3, no. 4, pp. e213–e216, Apr. 2021, doi: 10.1016/S2589-7500(21)00038-8.

[2] E. H. Shortliffe and M. S. Blois†, “The Computer Meets Medicine and Biology: Emergence of a Discipline,” in Springer, 2006, pp. 3–45.

[3] V. C. Kaggal 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.

[4] M. Khalil, P. Prinsloo, and S. Slade, “Realising the Potential of Learning Analytics,” in Online Learning Analytics, 2021, pp. 79–94.

[5] S. Shafqat, S. Kishwer, R. U. Rasool, J. Qadir, T. Amjad, and H. F. Ahmad, “Big data analytics enhanced healthcare systems: a review,” J. Supercomput., 2018, doi: 10.1007/s11227-017-2222-4.

[6] S. Shafqat, A. Abbasi, M. N. Ahmad Khan, M. A. Qureshi, T. Amjad, and H. F. Ahmad, “Context aware smarthealth cloud platform for medical diagnostics: Using standardized data model for healthcare analytics,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 7, pp. 299–310, 2018, doi: 10.14569/IJACSA.2018.090741.

[7] F. Shafqat, M. N. A. Khan, and S. Shafqat, “SmartHealth: IoT-Enabled Context-Aware 5G Ambient Cloud Platform,” in Studies in Computational Intelligence, vol. 933, Springer Science and Business Media Deutschland GmbH, 2021, pp. 43–67.

[8] M. L. Braunstein, Health Informatics on FHIR: How HL7’s New API is Transforming Healthcare. 2018.

[9] N. Shiri Harzevili and S. H. Alizadeh, “Mixture of latent multinomial naive Bayes classifier,” Appl. Soft Comput. J., vol. 69, pp. 516–527, Aug. 2018, doi: 10.1016/j.asoc.2018.04.020.

[10] M. S. R. Nalluri, K. Kannan, M. Manisha, and D. S. Roy, “Hybrid Disease Diagnosis Using Multiobjective Optimization with Evolutionary Parameter Optimization,” J. Healthc. Eng., vol. 2017, 2017, doi: 10.1155/2017/5907264.

[11] A. H. Osman and H. M. Aljahdali, “Diabetes Disease Diagnosis Method based on Feature Extraction using K-SVM,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 1, pp. 236–244, 2017.

[12] Z. Jia, X. Zeng, H. Duan, X. Lu, and H. Li, “A patient-similarity-based model for diagnostic prediction,” Int. J. Med. Inform., vol. 135, Mar. 2020, doi: 10.1016/j.ijmedinf.2019.104073.

[13] K. Ng, J. Sun, J. Hu, and F. Wang, “Personalized Predictive Modeling and Risk Factor Identification using Patient Similarity,” AMIA Summits Transl. Sci. Proc., vol. 2015, pp. 132–136, 2015.

[14] M. S. Ayhan, L. Kühlewein, G. Aliyeva, W. Inhoffen, F. Ziemssen, and P. Berens, “Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection,” Med. Image Anal., vol. 64, Aug. 2020, doi: 10.1016/j.media.2020.101724.

[15] G. Sathiya and P. Gayathri, “Automated detection of diabetic retinopathy using GLCM,” Int. J. Appl. Eng. Res., vol. 9, no. 22, pp. 7019–7027, 2014, Accessed: Oct. 10, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/pmc5961805/.

[16] T. Johnson, L. V. S. Lakshmanan, and R. T. Ng, “The 3W model and algebra for unified data mining,” in Proceedings of the 26th International Conference on Very Large Data Bases, VLDB’00, 2000, pp. 21–32, Accessed: Sep. 08, 2020. [Online]. Available: http://vldb.org/conf/2000/P021.pdf.

[17] I. D. Dinov, “Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data,” Gigascience, vol. 5, no. 1, p. 12, 2016, doi: 10.1186/s13742-016-0117-6.

[18] S. Shafqat et al., “Leveraging Deep Learning for Designing Healthcare Analytics Heuristic for Diagnostics,” Neural Process. Lett., pp. 1–27, Feb. 2021, doi: 10.1007/s11063-021-10425-w.

[19] J. C. Mandel, D. A. Kreda, K. D. Mandl, I. S. Kohane, and R. B. Ramoni, “SMART on FHIR: a standards-based, interoperable apps platform for electronic health records,” J. Am. Med. Informatics Assoc., vol. 23, no. 5, pp. 899–908, Sep. 2016, doi: 10.1093/JAMIA/OCV189.

[20] S. Shafqat, A. Abbasi, T. Amjad, and H. F. Ahmad, “Smarthealth simulation representing a hybrid architecture over cloud integrated with IoT: A modular approach,” in Advances in Intelligent Systems and Computing, 2019, vol. 887, pp. 445–460, doi: 10.1007/978-3-030-03405-4_31.

[21] J. Howard and S. Gugger, “Fastai: A layered api for deep learning,” Inf., vol. 11, no. 2, 2020, doi: 10.3390/info11020108.

[22] S. Shafqat, H. Majeed, Q. Javaid, and H. F. Ahmad, “Standard NER Tagging Scheme for Big Data Healthcare Analytics Built on Unified Medical Corpora,” J. Artif. Intell. Technol., Aug. 2022, doi: 10.37965/JAIT.2022.0127.