Open Peer Review on Qeios

# Rules Extraction, Diagnoses and Prognosis of Diabetes and its Comorbidities using Deep Learning Analytics with Semantics on Big Data

Sarah Shafqat[1], Zahid Anwar[2], Raihan Ur Rasool[3], Qaisar Javaid[1], Hafiz Farooq Ahmad[4]

1 International Islamic University, Islamabad
2 North Dakota State University
3 IBM Technologies
4 King Faisal University

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

## Abstract

Millions of people die because of diabetes each year. Furthermore, most adults living with this condition are juggling with one or more other major health concerns. These related diseases also known as comorbidities, coexist with the primary disease, but also stand as their own specific disease. The challenge that healthcare professionals face is that Diabetes Mellitus (DM) is difficult to differentiate into its six forms. This hinders timely and accurate diagnosis and proper treatment. This paper presents our research in developing a novel Artificial Intelligence (AI) based approach to analyze data of real patients having different comorbidity diseases for interpretation and finding inferences for diagnosis and prognosis of DM and its comorbidities in patients in different scenarios. Details are provided about the data models used, relevant feature sets and their association rule mining, deep learning analytical models developed, and results validation against various accuracy measures. The performance of several big data analytics platforms was validated for the different models for three different sizes of endocrine datasets with varying parameters. The data models were mapped to HL7 FHIR v4 schema that is flexible in adapting to diagnostic models for all diseases. Out of several analytical models evaluated, Louvain Mani-Hierarchical Fold Learning (LMHFL) was found to be the most promising in terms of efficiency and accurate explainable diagnosis through reflective visualizations of associated features.

**Sarah Shafqat**[1,a,*], **Zahid Anwar**[2,b], **Raihan Ur Rasoolg**[3,c], **Qaisar Javaide**[1,d], **Hafiz Farooq Ahmadh**[4,e]

[1]*Faculty of Computing and Information Technology (FOCIT), International Islamic University (IIU), Islamabad, Pakistan*

[2]*Department of Computer Science North Dakota State University (NDSU)*

[3]*IBM Technologies Australia*

[4]*Computer Science Department College of Computer Sciences and Information Technology (CCSIT), King Faisal University*

[a] *ORCID: 0000-0002-6080-6765, sarah.shafqat@gmail.com*

[b] *ORCID: 0000-0002-4608-4305, zahid.anwar@ndsu.edu*

[c] *ORCID: 0000-0001-9966-2466, raihan.rasool@ibm.com*

[d] *ORCID: 0000-0002-6827-0184, qaisar@iiu.edu.pk*

[e] *ORCID: 0000-0002-8545-9771, hfahmad@kfu.edu.sa*

[*]**Corresponding author**

**Highlights**

- This research contributes a great deal to the discipline of rules extraction for the right and extended feature set for DM diagnosis and associated comorbidity diseases using big data analytics platforms; RapidMiner and Orange Framework.
- The foremost advantage of this research is that the datasets are extracted from EHRs following a uniform data model developed on HL7 FHIR v4 schema that is flexible to adapt to diagnostic models for all diseases.
- Acquiring a real-time quality big EHR dataset of endocrine patients of considerable complexity for diagnosis was a challenge.
- Exhaustive exploration and validation of proposed DL heuristic models over open-source cloud platforms and frameworks were computationally sound and cost-effective.
- The proposed analytics are not a completely black box and designs are explained well with visualizations of results and performance.
- Proposed big data analytics with custom Deep Multi-Label Multinomial Learning experimented in Rapid Miner gave 100% accurate diagnostic results for records above 30k employing distributed HPC.
- But we found our proposed Fast-LMHFL model integrated with Fast.ai deep learning library for text classification in Orange framework optimum with visual clarity and explainable diagnostic results.

**Keywords:** Big Data, Neural Nets, Deep Learning, Healthcare Analytics, Diagnosis, Diabetes, Comorbidities, Endocrine Diseases.

## 1. Introduction

According to the World Health Organization, approximately 1.5M people worldwide died due to Diabetes Mellitus (DM), also commonly known as diabetes in 2019. It is estimated that there are currently 463M people with diabetes worldwide which will rise to 700M by 2045. DM is often associated with other endocrine system disorders that lead to the ten most common diseases such as; hypothyroidism, thyroid cancer, hypoglycemia, metabolic disorder, Addison's disease, Cushing's disease, polycystic ovary syndrome, thyroiditis, growth hormone problems and hyperthyroidism. Untreated or mismanaged DM may cause other chronic diseases external to the endocrine system, like; cardiovascular, kidney, liver, neuropathy and foot ulcers resulting in amputation. Furthermore, most adults living with this condition are juggling with one or more other major health concerns. These related diseases also known as comorbidities, coexist with the primary

disease, but also stand as their own specific disease. Just some of the many comorbidities identified in individual patient profiles in this paper include Premature Ejaculation (PME), Lower Back Pain (LBP), Transient Ischemic Attack (TIA), Benign Prostatic Hyperplasia (BPH), Lower Urinary Tract Symptoms (LUTS), acne, sleep starter (Jerk), Paresthesia and Throat Infection.

DM manifests in different forms known as type 1 diabetes, type 2 diabetes, gestational diabetes, latent autoimmune diabetes of adulthood (LADA), maturity-onset diabetes of the young (MODY) and neonatal diabetes mellitus (NDM). The challenge that healthcare professionals face is that DM is difficult to differentiate into its six forms. Since the comorbidity is associated with the DM type, this hinders timely and accurate diagnosis and proper treatment. Prognosis is different from diagnosis. The prognosis of any future disease occurrence may be inferred based on the profile of DM patients over time. For example, a prognosis may determine the condition of a DM patient after five or ten years based on his current and past conditions. Disease prognosis has its own challenges. Accurate monitoring and recording of patient condition with respect to time is important. Further if diagnosed together with co-existing diseases, the response to treatment would also help in forming the correct prognosis. In some cases, the results of our experiments showed comorbidity diseases did not match the diagnosis of particular patients but may be inferred as pointing to the formation of a disease in the future.

Artificial Intelligence techniques aim to allow machines to assist health care professionals in efficient and accurate diagnosis and prognosis of diseases. However, there are several challenges to this. Firstly, determining the correct features to be provided as input to the algorithm is of paramount importance. These features will follow traditional diagnostic steps undertaken by any clinician. Age, gender, vitals, symptoms, test results, final notes, and practitioner comments are significant inputs to the system for accurate inferences. Extracting the right feature set with correct values for diagnosis is quite challenging when there is still a lack of interest on doctors' part to enter this data digitally. Another challenge is to get access to these data features with values associated with anonymous patient profiles. Even with a dataset with the correct feature set, a challenge is to deal with incompleteness in many forms like constant dates of visits, missing feature values, misspelled and several forms of a single data value. The next step after cleaning and preprocessing this data is to study these features and their associations with their weightage on the desired results. When applying AI algorithms to develop any analytics, association rules mining is very important to find the relationship between features and weights attached to them for decision-making. Proper association rules extraction requires considerable effort, study and consultation with the doctors' team.

Electronic health records or EHRs are systematically collected and stored health information. They promise efficient data retrieval, improved care, a reduction in redundant tests, and faster diagnosis. The challenges to the efficient use of EHR technology for timely analytics and diagnosis of DM and its comorbidities are the lack of standardized data models, interoperability and data entry.

We have developed a methodology using Artificial Intelligence and machine learning techniques to address the above challenges. This paper presents our design of a uniform data model based on the HL7 FHIR v4 schema to store DM datasets. The model is flexible in that it adapts to all kinds of diseases not just DM. Three large datasets provided by Shifa International Hospital, Pakistan, of DM patients suffering from various DM comorbidity diseases were utilized for the

experiments. Natural Language Processing (NLP) a branch of Artificial Intelligence (AI) was used to extract data to populate the data models. Several deep-learning analytics models were developed to interpret and find inferences for the diagnosis and prognosis of DM and its comorbidities in patients in different scenarios. Due to the large size of the data sets two different big data analytics platforms; RapidMiner and Orange Framework were considered. A thorough validation of performance and accuracy was performed, and a discussion of the results is presented.

The novel contributions of this research are as follows:

1. This paper recognizes the comorbidities present in DM patients as a result of mismanagement or disease becoming chronic with age and the importance of analytics for automated diagnosis. The proposed methodology, as handled by experts, has been codified in this paper. It details every stage of analytics design and architecture, from data modeling to data extraction, and analyzes it to diagnose DM and comorbidity diseases.
2. The challenge of providing a quality big dataset of considerable complexity to analyze for diagnosis was catered to in this paper. Three large datasets of real endocrine patient data were utilized for evaluation. The patients were suffering from different types of DM with multiple diagnoses from multiple visits. The first two datasets comprised data from 100 patients each and the third related to 14407 patients. The cloud platforms and frameworks explored for managing large datasets were found to be computationally sound and cost-effective for analyzing this intensive data.
3. These datasets are created using mapping onto a uniform data model developed on HL7 FHIR v4 schema that is flexible in adapting to a diagnostic model for all kinds of diseases. Furthermore, data extraction challenges for populating data models using NLP for meaningful analytics are addressed along with a discussion of how the considered datasets were prepared in ICD-10-CM for providing uniformity in medical nomenclature for understandability on the international front.
4. Analytical models utilizing AI-based deep learning heuristics were developed. Current research in this area faces challenges of algorithmic bias and uncertainty. These challenges have been adequately addressed. Algorithmic bias is addressed by regularizing the analytics designed not to be deterministic and over-fitted. The question of uncertainty is resolved with probabilistic results and flexible weights given to features for the best results and keeping room for expert input.
5. Analytics were checked for the optimum solution and parameters setting. The results demonstrated very high accuracy. The best results were obtained within Deep Multi-Label Multinomial Learning in Rapid Miner. The other case discussed the design of the Louvain Mani-Hierarchical Fold Learning (LMHFL) algorithm in the Orange framework. In this case, the fast.ai deep learning library was used for text classification for the best heuristic design of the LMHFL.

The authors of this paper followed a rigorous approach to prepare different-sized datasets from a standardized endocrine data model prepared on FHIR v4.0 HL7 schema with endocrine diseases labeled with ICD-10-CM codes. These datasets were preprocessed at many stages during analysis for best predictions for the diagnosis of DM and its comorbidities.

The paper is structured in nine sections for readability. Section 1 is allotted to the introduction. In section 2, previous analytics that were applied to diabetes datasets and their comorbidities are studied. Section 3 elaborates on the methodology applied for experimentation in the paper. Section 4 then explains the preprocessing and data modeling at

length. In Section 5, researchers realized that datasets needed further cleaning and pruning for better diagnostic accuracy during experiments. Section 6 sheds light on the design and architecture of proposed analytics models. Section 7 elaborates on the experimental studies and results gathered. Section 8 discusses the insights from further analysis of these experimental results. Section 9 evaluated different analytical algorithms and cloud analytics frameworks and models designed for diagnosis. Section 10 focused on analytics results of individual patients selecting some particular patients' profiles for analysis. Section 11 and 12 shed light on the Fast-LMHFL model applied on tabular and text data in different settings for diagnosis and prognosis of comorbidities of individual DM patients. Finally, section 13 explains the findings and concludes with a future vision to fine-tune these analytics.

## 2. Related Work

Our research benefits from previous research done on DM, EHR, AI applications in health care, comorbidity diseases, ANN, and supervised, semi-supervised and unsupervised learning. This section provides coverage of these areas accordingly.

DM is often understood to be misdiagnosed or commonly diagnosed as Type 1 or 2 DM where there are other six forms known as gestational diabetes in pregnant women, latent autoimmune diabetes in adulthood (LADA) [1] also considered as type 1.5 diabetes. LADA is nearer to type 1 diabetes but is often misdiagnosed as type 2. The same is the case with maturity-onset diabetes of the young (MODY) [2] and is often misdiagnosed as type 1 or 2 diabetes. Similarly, Neonatal Diabetes Mellitus (NDM) [3] is often misdiagnosed as type 1 DM. Some of these rare forms of DM like; gestational or T2D are present in our datasets with other comorbidity diseases and pre-diabetic cases. In both [4], [5], it is mentioned that the people suffering from type 2 diabetes (T2D) are more, therefore, researchers investigated its causes of occurrence based on lifestyle, family linkage, food consumption, age and other medical complications. DM is best managed by controlling hyperglycemia and other complications to occur [1]. The analytical work has been done using various data mining tools on available datasets, but accuracy is not guaranteed. Gestational diabetes in early pregnancy was analyzed using 33935 records from the study cohort in West China Secondary Hospital [6]. Conceptual design [5] is given for implementing a data mining method with improved accuracy proposing a self-organizing map (SOM). The diabetes dataset in [5] is taken from the UCI ML repository. The evaluation was done on experimental results and SOM was found better in comparison with random forest, Naïve Bayes, decision tree, MLP or SVM. A case study with a reflection of diabetes big data as in Table 1 of [7] was conducted on data collected by the Diabetes Screening Complications Research Initiative (DiScRi) of a regional Australian university. Figure 3 illustrated in [7] gives a glimpse of how to tackle the complexities of big data analytics. Early detection of Diabetes Mellitus (DM) is important [8] to safeguard patients from its chronicity and other complications that may fail other organs like; the kidney, eyes, heart, nerves and veins. A generalized way to select optimal features was established for classification and here it reflects that specificity got from decision tree and random forest is higher where Naïve Bayes gives an accuracy of 82.30%. DM is termed globally as a fatal disease [9]. Interpretation of diabetes data from patterns to diagnose in an efficient way using data mining and ML techniques is therefore found crucial.

Researchers presented previous studies carried out by taking Electronic Health Records (EHRs) and applying analytics to

them. Below are subsections on EHRs, ML and AI applications in healthcare, co-occurring or comorbidity diseases attached with DM, diabetes analytics, ANNs or supervised learning, semi-supervised learning and unsupervised learning with reinforcement discuss previous research in detail. Major limitations found were incompleteness of feature set, size of data in terms of records, comorbidity diseases undertaken, etc., uniformity in the data model and its mapping on FHIR HL7 standards and ICD-10-CM codes for the diagnosis of diabetes and its comorbidities, the generalizability and interoperability was missing with respect to the universal diagnostic model. Complete information on feature sets and 'Diagnosed' classes within datasets is given in detail in Tables 1 and 2.

## 2.1. Electronic Health Records

Electronic health records or EHRs [10] are systematically collected and store health information. EHRs have improved access to patients' medical profiles including symptoms, lab test results, diagnosis and prescribed medicines and allergies formed with demographics. Added advantages of EHR are efficient data retrieval, faster communication of quality indicators to hospital administration and reduced costs of health management and insurance benefits. Overall, this allows for improved care in the form of a reduction in redundant tests, faster diagnosis, and reduced errors in drug dosage and allergies.

## 2.2. Machine Learning and Artificial Intelligence Applications in Health Care

Machine Learning (ML) and Artificial Intelligence (AI) have revolutionized the way to extract data from EHRs for analysis and Natural Language Processing (NLP) is commonly used for text analysis in clinical notes. There are several companies working to analyze healthcare big data mentioned in Table 2 of [10]. The dissertation [11] reflects in detail on the usefulness of healthcare analytics in solving problems and adding value to preventive care. Innovation in analytics [11] is thus required for understanding complex data in healthcare. Essay 1 of this study focused on feature engineering and disease co-occurrence relationship networks. Essay 2 proposed an analytics method to identify patterns in missing and incomplete data which would train multiple reduced models minimizing imputation and missing values. Essay 3 proposes sensor-based analytics for managing sound levels in the workplace. A recent study is going on to integrate a deep learning model with transfer learning approaches for clinical text analysis that promises improved performance [12]. Transfer learning is known for modeling ML algorithms for NLP tasks therefore, it may also be used to model deep learning frameworks as depicted in figure 2 of [12]. The paper [13] again puts emphasis on the importance of feature selection fitting in large datasets to avoid over-fitting for the application of ML. There are some algorithms with automated feature selection. For novelty in feature selection the stability selection (SS) approach [13] is introduced using a genetic algorithm (GA) iteratively on subsamples of features in records. The fitness of genetic or evolutionary algorithms is validated through the area under the curve (AUC) by continuous optimization using known methods like particle swarm optimization (PSO) or hyper-parameter optimization [14]. The top 4 SS features using GA improved AUC results. This analysis was done on the nationwide inpatient sample (NIS) database with the support of the European Cooperation in Science and Technology (COST).

## 2.3. Comorbidities

Comorbidity diseases are those that co-occur with a primary disease like in our case it is diabetes mellitus. Predicting complications [15], [16] associated with chronic diseases like; diabetes, is required to make personalized treatment plans for patients. Obesity and metabolic disorders are seen in patients suffering from DM leading to other chronic illnesses like; thyroiditis, thyroid cancer [17], hyperthyroidism, Addison's disease, Cushing's disease, growth hormone problems[18], polycystic ovary syndrome [19] or liver cirrhosis [20] making DM into a fatal disease if not managed[21]. The impact on patients with Liver Cirrhosis having DM is studied in [20] to know the chances of complications faced by pre-cirrhotic and post-cirrhotic patients. Group A and Group B were statistically analyzed for cirrhosis along DM and cirrhosis without DM respectively. There were 116 patients suffering from DM and cirrhosis out of which 59 developed Hepatogenous Diabetes (HD) and 48 developed Antecedent Diabetes (AD). Cirrhosis patients with or without DM were not much differentiable in terms of complications shown in Table/Figure 1 of [20] even later it was proved that cirrhotic patients with DM were more prone to complications. The study continued by dividing Group A patients into subgroups {Antecedent Diabetes (AD)}: having patients diagnosed with DM prior to diagnosis of cirrhosis and {Hepatogenous Diabetes (HD)}: patients diagnosed with cirrhosis prior to the diagnosis of DM. Through various statistical inferences it was seen that diabetic patient is disposed to serious liver diseases and DM greatly associates with higher complications and mortality in patients forming cirrhosis [22]. Risk occurrences [15] in type 2 diabetes (T2D) on diagnosis were studied through longitudinal medical records of patients. Relationships that are captured include a) between risks from complications arising from T2D, b) among different risk factors, and c) between selection patterns of risk factors. The proposed method [15], is of hierarchical Bayesian framework that incorporates domain knowledge in high-dimensional data based on an informative subset identified from coefficient shrinkage. This method is found significant for identifying patterns of associations in risks to give clinical insights by using it in state-of-the-art healthcare applications. This work is limited by taking temporal information into consideration and opens a path for better feature representation for predicting risks. It also suggests that as T2D, different correlations may be found with other severities of DM. The addition of domain knowledge of risk factors would further enhance the performance of the given method [15]. Comorbidities associated with T2D [23] are critical to predict to save society from the chronicity of this disease. The nationwide data from hospitals and medical prescriptions of more than two lakh newly diagnosed T2D patients went through ML application using logistic regression, random forest and gradient boosting models (Figure 2). This mechanism predicts 5yrs risk of heart failure (HF), chronic kidney disease (CKD), myocardial infarction (MI) and cardiovascular disease (CVD). Danish nationwide internationally classified (ICD-10) patient registers in Denmark, with respect to hospitalization, diagnosis, prescription of drugs, etc. are found better in producing results than reference models [23]. The correlated features in registers highly signified the challenges lying in the interpretation of feature importance. Therefore, additional models with extended parameters using other frameworks and performance metrics are awaited [23]. Figure 1 below shows the extended features list of diagnoses of DM and other comorbidity diseases from our dataset that is extracted for analysis.

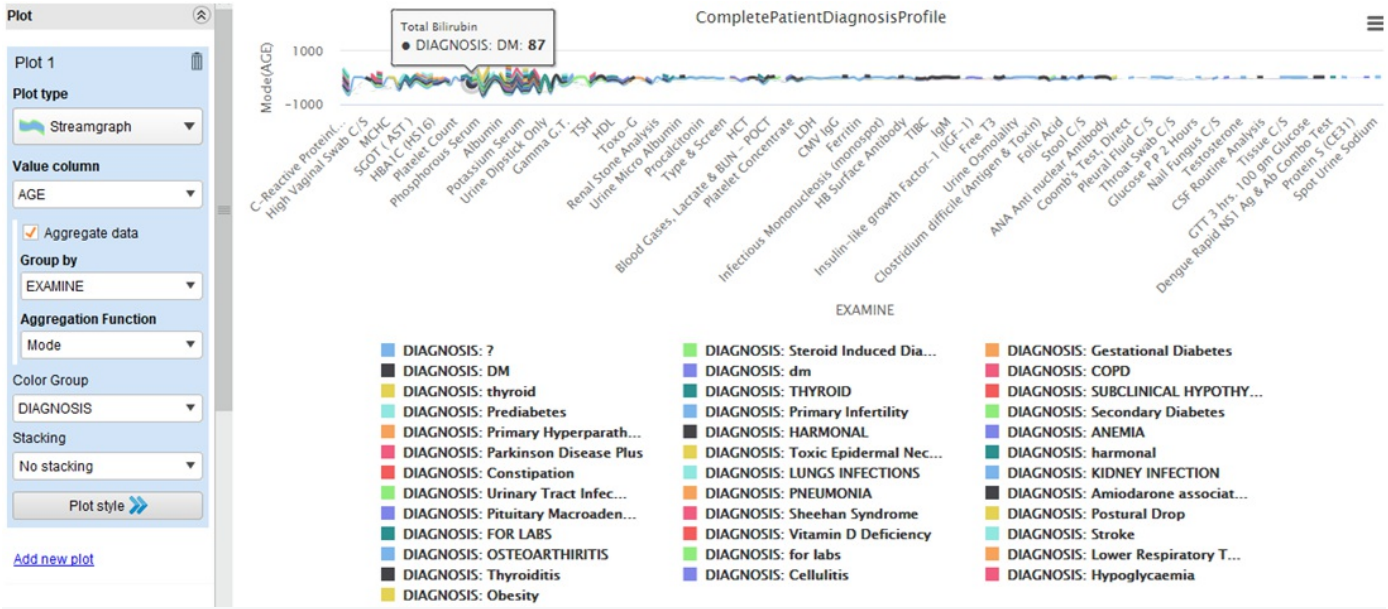



**Figure 1.** RapidMiner Stream Graph showing Diagnosis of DM and its 30 comorbidity diseases related to age and corresponding examination

## 2.4. Diabetes Analytics

Detailed study [4] is done to investigate multi-layer ML algorithms for improving analytics models for diabetes. Analytics performs best on accurate data and meaningful structured information that is made available by medical domain experts for supervised, semi-supervised or unsupervised learning [24]. A detailed ontology of diabetes[25], therefore, is looked into to get an insight into the dataset required to put through analytics for finding associations to diabetes and its comorbidity diseases [21] in the endocrine system and outside of it. Hence, our diabetes mellitus dataset of 14407 endocrine patients over a span of 2 years; 2018 and 2019 counted 30 multiple diagnoses of associated comorbidity diseases (omitting DM itself and undiagnosed cases in Table 2 of section 5) is well elaborated in Figure 1 of section 2.

Prevalence of Type 1/Type 2 Diabetes is also found in major cities of Saudi Arabia and as usual in females, its occurrence is higher. It analyzed some key features like; Body Mass Index (BMI), tests; HbA1c, Impaired Fasting Glucose (IFG) and others (as in Table 1.4 of [4]) with results to reach a diagnosis of diabetes, pre-diabetes and identification of its causes. The abnormal blood glucose (BG) levels result from the occurrence of DM caused by metabolic disorders and become a reason for major complications and increased mortality if not managed well [26].

## 2.5. Artificial Neural Networks (ANNs) or Supervised Learning

ANNs become the foundation of several ML techniques applied for the prediction of DM and are compared for the level of accuracy achieved. The study in [26], mainly worked on the prediction of BG by pooling similar patient profiles to determine an alarm system for any criticality about to occur in type 1 DM. The known algorithms are elaborated in Figure 2 and were referred to in [26] specifically. ANN, SVM with kernel and Gaussian method[14], genetic or evolutionary algorithms (EA) that are biologically inspired for learning from events and use genetic programming to evolve, random forests are also known for an ensemble of trees approach to mine the mode or mean of the probable class. ANNs are

designed for labeled and unlabeled data moving from supervised to semi-supervised and unsupervised methods or advanced neural networks as seen in Figure 2.

## 2.6. Semi-Supervised Learning using Labeled or Unlabeled Data

Analytics are seen to give better results with hybridization as support vector regression was combined with random forest regression or different features based on its weights were nested to better predict BG levels by using; glucose profile, meal-derived glucose, energy expenditure routine and plasma insulin level. The accuracy of results was determined using recall and precision in several clinical studies. In their work [4], researchers found Naïve Bayes outperforming the other three supervised ML techniques; decision tree, neural networks and support vector machine (SVM) for analyzing large-scale health data. Still, the flexibility to handle complexity and non-linearity in data allows the researchers to select deep learning, multilayer perceptron (MLP), and SVM as in [27][28][29]. To conclude, researchers [4] prepared a generalized confusion matrix for selection from different classes of algorithms. There is still a chance of misinterpretation and false perceived accuracy, therefore, F-Measure is considered. Ensemble or Hybrid Modeling is also considered for enhancing accuracy by combining two or more algorithms. The dataset did not have clinical notes; therefore, NLP was not applied. In another study [5], the need for diagnostics for diabetics is felt with its rise in the global population. The study[9] was done on the PIMA Indian Database by employing Weka to perform mining to diagnose DM. Bootstrapping was done with resampling using Naïve Bayes, KNN, and Decision Tree to achieve increased accuracy. Diagnosing diabetes at the initial stage through data mining is surveyed [30]. ‘CoLe’ is proposed as a multi-agent having multiple data miners for higher accuracy. Apriori as associative rule mining was applied to create equal interval bins for continuous variables to classify diabetes. Estimation Maximization (EM) with ID3 used as a hybrid prediction model for diabetes classification gave 91.32% accuracy. Another expert system for the diagnosis of DM with an extended learned classifier achieved a greater accuracy of 91.3% by simply using if-else rules. SVM with Naïve Bayes proved to give an accuracy of 97.6% for DM prognosis complimenting the results from the ensemble model proposed in [4] on a dataset of 768 instances and a single class. Likewise, other algorithms; C4.5, J48 (decision tree), KNN, MLP and ANFIS also emerge as proven for their near-to-accurate results and the performance increases when combined [30]. Care for diabetics accounts for 12% of health expenditure globally [31] with an estimation of 425 million population affected by it. Embedding applications with advanced AI for caring for persons with diabetes (PWD), caregivers; clinicians, family, nurses, and pharmacies is felt promising. Researchers queried PubMed for the terms; ‘diabetes’ and ‘artificial intelligence’ publicly available, excluding the technicality and retrieved 450 articles. Transformation in diabetic care [31] is perceived in four areas; i) auto-retinal screening, ii) prediction of risk-stratified population, iii) clinical decision support system, and iv) self-managed patient care. This transformation is seen nearer with AI apps coming into the market with the promise to monitor glucose control and reduce hypoglycemic episodes, preventing complications and comorbidities due to mismanagement of diabetes. The clinical decision support system again underwent a study on top of T2D risk models [32] that included only 90 respondents of which 5 to 17 participated in each experimental round. Two sets of data were defined as T2D screening and T2D care. In the user scenario, both studies consider the same priorities with individual care provided by doctors. Management on the other hand tends to see it at the population level for the provision of needed tools and guaranteed satisfaction. For individual satisfaction visual analytics [32] proved to focus attention on events and parameters in attractive settings but

complexity could not be ignored with perceived potential.

Recently it has been witnessed in [33] that deep neural networks (DNNs) basically enhance the capability of any machine learning algorithm by adding multiple neural network layers and statistical measures to; CNN, Bayesian networks [34], MLP [28] or as common AI algorithms like; K-Means[29], ANN or decision tree to give results with vector and complex datasets that initially worked on scalar datasets only. The Deep Learning Algorithmic tree in Figure 2 represents a clear picture for enhancing the capability of ML and AI algorithms by means of different statistical measures discussed here are; Binomial, Bernoulli, Gaussian and Multinomial distribution. In Figure 2 researchers elaborate how semi-supervised learning takes labeled and unlabeled data both into consideration and with hybridization of supervised and unsupervised algorithms often give semi-supervised algorithms for best analytical results.

## 2.7. Unsupervised or Reinforcement Learning

Researchers are finding DNNs very helpful to diagnose patients using medical EHRs with precision equaling to qualified clinicians as in [33] diabetic retinopathy (DR) is under consideration. The near-to-accurate DNNs do have to consider the possible uncertainties and not lead to overconfident diagnoses risking the patients' health. Therefore, there has to be room for flexibility of selection and weights given to feature sets and expert inputs at all stages of diagnosis for reinforced learning [35]. Paper [33] puts emphasis on doctors finding it difficult to explain or filter the risk of uncertainty in diagnosis. Bayesian DNNs are found computationally expensive and trade performance to predict the uncertainties in diagnosis using dropout, regularizers or batch normalization. The alternative to Bayesian [34] is ensemble DNNs where each node is initialized at random to sample diverse accurate predictions improving the single network performance but it costs the training and interpretation. Building on current research [33], experimentation is done using test-time data augmentation (TTAUG) in DNNs to be intuitive through a data-driven approach. The diagnostic uncertainty was validated by matching clinicians' differences of opinions with the rate of uncertainty predicted in diagnosis proportionally. Two datasets were taken with retina images from the Kaggle Competition and the Indian Diabetic Retinopathy Image Dataset (IDRiD) for analysis on a severity scale set by International Clinical Diabetic Retinopathy. Generalization is either achieved through patient similarity-based model [36][37] or performance was checked in [33] for the whole network for uncertainty distributions up to five levels of the DR scale. The evaluation was done for two classes of DR diagnosis; mild and moderate. CNN with Residual Networks (ResNet) was applied on these DR image datasets [38] modified to fully connect all layers with an additional layer before softmax. It used parametric rectifier logical units (PReLUs) as an activation function on the 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 50,000 iterations starting with 20 images in a mini-batch. L2 regularizer was used in the first stack of layers that adopted the L2 regularization measure in the fully connected layer. The network adjusted itself to the distribution of classes. Data augmentation of 0.9 was done at every iterated cycle that resulted in 11350000 images out of which 10215000 were generated randomly. Data augmentation measure was not found enough for predicting uncertainty estimation and TTAUG was proposed for deterministic classes. ResNets were modified to act as

DNNs and results were gathered for evaluation using TTAUG and validation through experts' feedback.

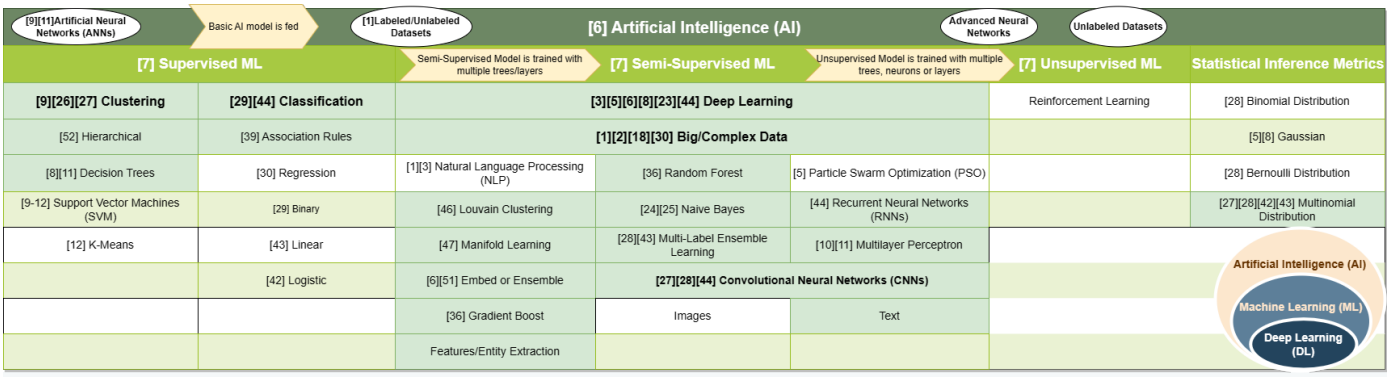



**Figure 2.** Deep Learning Algorithmic capability Tree for different variations in Statistical measures named here as Multinomial, Gaussian, Bernoulli and Binomial Distribution applied on ML and AI known algorithms.

Finally, ensemble predictions [33] were found more robust even if lacked in evaluating the uncertainties in DR diagnosis hence, it 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 under-confident results yielded through the ensemble were found to be highly accurate. Automation simplifies and speeds up the design and experimentation to validate the performance and accuracy achieved from deep learning for the diagnosis of DR [38]. In this paper [38], CNNs were explored with multinomial classification models to increase the sensitivity of the algorithm to misclassification of disease. Two datasets were taken; one from Kaggle of 35000 images with five classes; normal, mild, moderate, severe and end stage, where the other dataset was Messidor-I with 1200 colored images with four classes omitting 'the end stage' label as stated in another dataset. The Kaggle dataset is recognized as unclean with false labeling and poor-quality images. The model here was first trained on the Kaggle Dataset and then on a smaller but more reliable dataset of quality images. Finally, 550 images were interpreted and validated by ophthalmologists. Transfer learning made it possible to use GoogleNet and AlexNet as a foundation for binary, tertiary and quaternary classification. The final model [38] used a rapid neural networks prototype on deep learning interactive framework DIGITS powered by TensorFlow for training. Multi-label classification on the Kaggle dataset was found limited as sensitivity for binary or binomial classification was 95% with poor sensitivity of 7% for mild DR as shown in Table 2 of [38]. Messidor dataset, however, showed a more accurate measure for multi-label classification as no DR or severe DR were 85% and 75% sensitive with mild DR of 29% sensitivity. Transfer learning then expedited the gain to sensitivity for mild DR with 17%. It is seen that multinomial distribution is often chosen where multi-label classification is needed [39]. The multinomial distribution is an efficient method to generalize the binomial distribution for success and failure by expanding it to multiple graded scales from low to extreme scenarios. Like in [38], it was no DR or severe DR but within there were different independent stages of DR; mild DR and moderate DR to which higher sensitivity was required by the classification model. Generalizing a model for superior results is also not ignored and the Sparse Bayesian Extreme Learning Machine (SBELM) is developed on top of the extreme learning machine (ELM) for sparse, generalized and quicker results generation. SBELM has 3 layered neural networks with the probabilistic model using Bernoulli distribution for binary classification and went on to do multi-label classification through pairwise

coupling by binary classifiers [39]. Uncertainty constraint was of concern here as well, therefore, multinomial distribution was employed for unambiguous multiclass distribution. Hence, multinomial Bayesian extreme learning machine (MBELM) was proposed in [39] for 5% better results in comparison to SBELM in test accuracy and 94 times smaller model size as previously it is seen that model size exceeded with increased classes. MBELM [39] solved the problem of probabilistic distributed classes determining the most likely class among the distribution that was lacking in ELM and SBELM. Laplace was used to increase the likelihood of probable class and the softmax activation function was used replacing the sigmoid function that was used in SBELM for binomial classification.

## 3. Methodology

This paper devises a methodology to experiment with auto and custom deep learning analytical models designed and trained on cloud analytical platforms and employing chosen resources in different settings. The methodology as illustrated in Figure 3 is detailed below.

1. The first step was devising a data model on which the datasets provided can be mapped. The data model had to be standardized for interoperability using HL7 FHIR v4.
    1. The desired dataset was then requested based on the entity-relationship model (ERD) designed based on universal features extracted from the FHIR v4 schema.
    2. The generalizable diagnostic dataset for DM and its comorbidities was then given in stages.
2. Features relevant to patient profiles and diseases, in particular, were then studied to understand associative relationships in traditional diagnosis.
    1. The datasets were preprocessed and cleaned for missing, misspelled or multiple forms of entries for the same inputs and carefully labeled with diagnosed diseases and ICD-10-CM codes.
    2. The uniform features set were then finalized during experimentation.
3. This selected features set became the input to transform a traditional diagnostic model to analytics using deep learning heuristics.
    1. Previous work was rigorously studied to find limitations and gaps.
    2. A suitable set of analytical cloud platforms was chosen based on the previous study.
4. Finally, in different experimental settings the proposed analytical heuristics were tested and validated.
    1. Results were analyzed.
    2. The diagnostic rule sets were found through understanding inferences gained from visual representations and accuracy results.

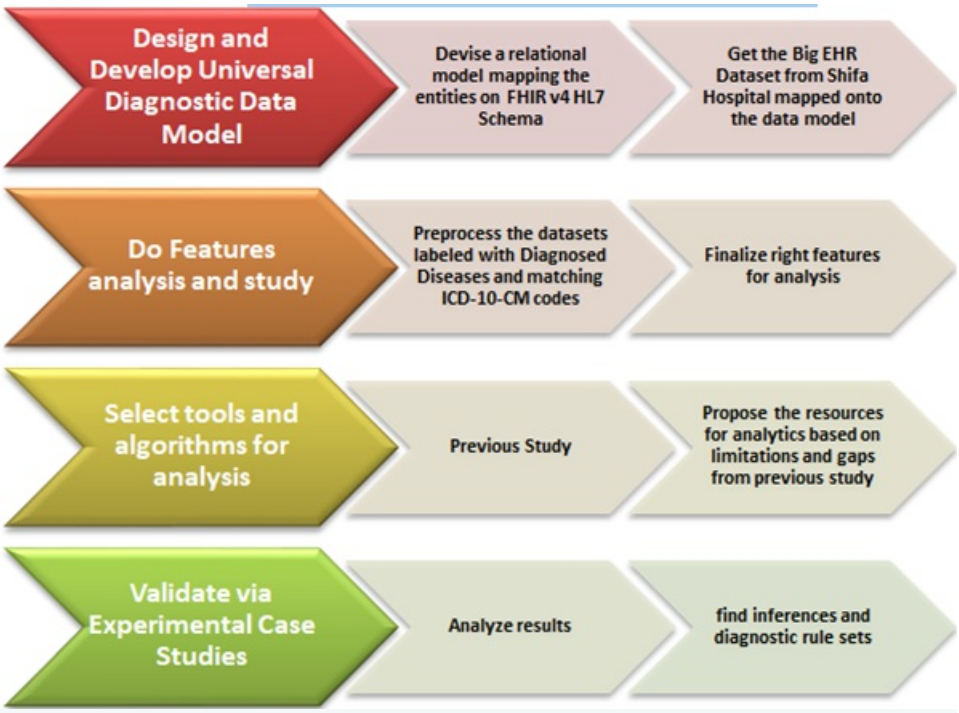



**Figure 3.** Methodology selected for experimentation and formulation of Big Data Analytics for Diagnosis of DM and its comorbidities.

Open-source cloud platforms that were studied for this research were Qlik Sense, Rapid Miner and Orange Framework on Anaconda integrated with Fastai deep learning library with others that were available by default.

Three scenarios are considered for experimentation:

1. Running analytics on complete data sets of multiple patients to find associations between patients for similarity in features and the relationships between different endocrine diseases and particularly with DM.

The features set belonging to the data model is as follows:

$$F_{k[rows=2844\rightarrow 33185:cols=3\rightarrow 18]}\varepsilon D_{[j=5\rightarrow 66]}\varepsilon P_{[i=100\rightarrow 14407]}, \qquad (1)$$

where features $F_k$ are related to form single diagnosis $D_j$ that further belong to each patient $P_i$ in a complete dataset $\Theta$

$$Patients(P_i) \rightarrow Set\Theta:[P_0, P_1, P_2, \ldots P_i] \qquad (2)$$

$$MultipleDiagnosis(D_j) \rightarrow SetP_i:[D_0, D_1, \ldots D_j] \qquad (3)$$

$$\text{Features for Diagnosis}\left(F_{k[rows:cols]}\right) \rightarrow \text{Set}D_j:[\text{rows}_r^{33185}: \text{cols[gender, age, note, exam, test, result, PC,...]]}$$

2. Further these datasets were then analyzed and studied for different diseases suffered by individual patients and specifically DM patients.
3. Finally, for better analysis and accuracy the text fields;

$$T_{m[rows:cols[Note,PC]]}$$

has Note and PC (representing Practitioner Comments), were separated from the tabular features to run NLP technique

through Fastai v2 library.

A feature set is now understood as:

$$T_{m[rows:cols[Note,PC]]} \varepsilon F_{k[rows_r^{33185}:cols[gender,age,note,exam,test,result,PC,...]]} \varepsilon D_j \varepsilon P_i$$

where text features are denoted as $T_{m[rows:cols[Note,PC]]}$

belonging to features $F_{k[rows:cols]}$ that are related to form a single diagnosis $D_j$

that further belongs to each patient $P_i$ in a complete dataset $\Theta$

Likewise, there were some auto and custom ML and deep learning models designed, tested and validated against various accuracy measures.

1. Firstly, some common and known machine learning models were tried using Jupyter Notebook and Python in section 9.
2. Secondly, when limitations were recognized in high-performance computing we shifted to open-source cloud platforms for big data analytics.

## 4. Data Preprocessing and Modeling

The datasets taken were three for variant numbers of DM and endocrine patients with multiple diagnoses from multiple visits as described in Tables 1 and 2. The first two datasets are of 100 patients provided synchronously to develop and complete the required features list. The first dataset is complete in features but is limited to diseases; DM, Thyroid and Hormonal. This dataset consisted of 3650 records when converted to a metadata sheet for analysis mapped on HL7 FHIR v4 schema. This dataset was then coded in ICD-10-CM for uniformity in medical nomenclature for understandability on the international front. The second dataset was of another set of 100 DM patients diagnosed with comorbidity diseases during multiple visits (See Table 2 in Section 5). This dataset was mapped onto the same features list that was set on the prior metadata sheet made on HL7 FHIR schema to keep our data model uniform to integrate on SmartHealth cloud [40] for future implementation. The third dataset considered was of 14407 endocrine patients with several diseases as is seen in Table 2 of Section 5. Each patient suffered from a unique endocrine or related disease and associations for these diseases in a single patient, therefore, were not identifiable.

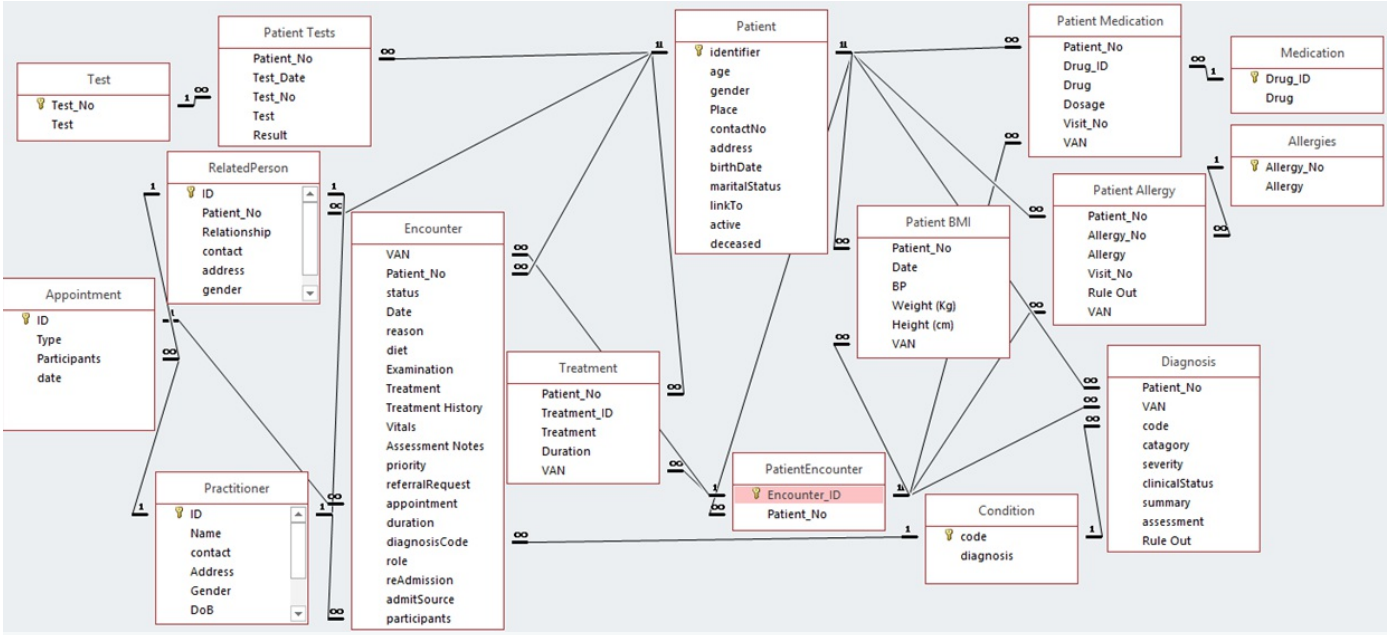



**Figure 4.** Entity Relationship Diagram (ERD) for Relational Data Model for Patient Profiles

**Table 1.** Features List selected in three Data Sets.

| | Dataset 1: 3650 instances of 100 patients | Dataset 2: 15696 instances of 100 patients | Dataset 3: 87803 instances of 14407 patients |
|---|---|---|---|
| **Features** | PatientID, Age, Gender, VAN, Appointments, Note, Test Date, Examination, Test, Result, Assessment, PC, Diagnosis, ICD-10-CM | PatientID, Gender, Age, VAN, Appointments, Note, Test_Date, Examination, Test, Result, Assessment, PC, Diagnosis, ICD-10-CM | PATIENT_ID, VAN, VISIT_DATE, AGE, GENDER, EXAMINE, TEST, RESULT, ALLERGY, NOTE, ASSESSMENT, PC, DIAGNOSIS, ICD_CODE, MEDICINE, STRENGTH, UNIT_DESCRIPTION, DAYS |

Three Metadata sheets; Dataset 1, Dataset 2 and Dataset 3 in Table 1 are taken for our experiments that are derived from separate entity records normalized in MS Access shown in Figure 4. Three metadata views of 100 and 14407 endocrine patients are created from it on HL7 FHIR v4 standard having at most eighteen features; PATIENT_ID, VAN (Visit Account No.), VISIT_DATE, AGE, GENDER, EXAMINE, TEST, RESULT, ALLERGY, NOTE, ASSESSMENT, PC (Practitioner's Comment), DIAGNOSIS, ICD_CODE, MEDICINE, STRENGTH, UNIT_DESCRIPTION, DAYS. Our standard data model however has other features as well that are not taken in this study. It was because the data provided by Shifa International Hospital was limited and did not hold all the features given in the data model. Features like RelatedPerson and PatientEncounter were to know the medical issues associated with the family history of the patient and the number of patient visits to the hospital.

The diagnosis is made on some primary features extracted from the data model developed, shown in Figure 4, and that is 'Examination' as listed in the entity 'Encounter', 'Test' is the field in entity 'Patient Tests' (list can be seen in Table 3) recommended by the doctor based on 'Assessment Notes' taken from patient's 'Encounter' visit. The lab 'Result' produced helps in making practitioner comments (PC) and diagnosis listed in the 'Diagnosis' table as a summary and diagnosis code.

All these relational tables as received from Shifa International Hospital MIS Department were de-normalized into metadata sheets taken here as Dataset 1, Dataset 2 and Dataset 3. The feature sets are detailed in Table 1.

The International Classification of Diseases (ICD) and the associated health issues list maintained by the WHO contains codes for diseases, their symptoms, abnormalities in health conditions and societal problems and all causes of injuries and disease. ICD-10 came into being in 1983 and was endorsed in 1990 by WHO to be used by member states from 1994 onwards. Currently, work is going on its next version that would become public in 2022 as ICD-11. ICD-10-CM includes many new diagnostic codes compared to the old version, ICD-9 and has over 70,000 codes. ICD-10 also has procedural codes but those are not related to this research.

These Metadata sheets have rows labeled with respective ‘Diagnosis’ with the corresponding ‘ICD-10-CM’ code from (icd10data.com) entered manually. This data is fed into our analytical cloud platforms to get explainable results for supervised, semi-supervised and unsupervised learning [24] discussed in section 5 to extract association rules for analysis [41].

Each row in our datasets is an instance of a patient diagnosis considered as a node having features as in Figure 5 and 6. Our deep learning analytics clusters all patient nodes with respect to the diagnosis as seen in Figure 7, 12, 13 or 14. These nodes are positioned as per the feature set. Custom feature sets as part of eight rules were mostly tried. PatientID or Patient MR no. are the same and are selected to find an inference particular to patients. In Figure 15, Principal Component Analysis (PCA) shows the associations of DM patients with respect to age and gender. PCA statistically shows that DM is found in later ages mostly in females. Lab tests with the corresponding results are major factors in diagnosing these patients. Deep learning is a self-learning mechanism through given data but if features are classified and given weights as for the set biomarkers the resultant inferences are quicker to get with maximum accuracy. Clinical notes and practitioner comments are also important to review symptoms and previously prescribed medicines, which can expedite the diagnosis of DM and the identification of any comorbidities or associated risks. Visit Account No. (VAN) may show us the number of visits and diagnoses one patient has gone through. Appointment and test dates also depict temporal information about patient visits and lab tests but in our datasets this information is limited. Examination is another feature that groups certain tests into one category and forms a relation to the identification of a particular diagnosis. In other cases, it is also possible that we first predict recommended tests for given clinical notes and then based on the results associated with these tests and practitioner comments further predict the corresponding diagnosis class. Finally, we may consider diabetic patients’ profiles for finding comorbidity diseases as well as in our third dataset of 14407 DM patients, each having a single diagnosis. Most common rules extracted from our datasets, as depicted in Figures 5 and 6, are explained and defined below:

- Rule 1: Patient MR No. -> Multiple Diagnoses of Endocrine Diseases -> Diabetes (class) -> Comorbidities (class) (Figure 22 and 29)
- Rule 2: Age - Gender -> Diagnosis of DM (class) -> Comorbidities (class) (Figure 7, 15, 16 and 17)
- Rule 3: Tests -> Results -> Diagnosis (class) (Identification on multiple classes of diagnosis) (Figure 7, 13 and 14)
- Rule 4: Tests -> Results -> Ranges (Biomarkers for normal, mild risk and high risk) -> Diagnosis (class) (Identification

on multiple classes of diagnosis) (Figure 18 and 19)

- Rule 5: Notes -> Practitioner Comments -> Symptoms and prescribed drugs -> Diagnosis of DM -> Complications -> Comorbidities (Figure 18 and 19)
- Rule 6: PatientID, Gender, Age, VAN, Appointment, Note, Test Date, Examination, Test, Result, Assessment, PC -> Diagnosis (class) (Identification on multiple classes of diagnosis) (Figure 22)
- Rule 7:

  a. PatientID, Gender (Label), Age, VAN, Appointment, Note, Test Date, Examination -> Test (Prediction)
  b. Test (Predicted), Result, Assessment, PC -> Diagnosis (Cluster) (Figure 26 and 27)

- Rule 8: PatientID, Visit Date, Test Date, Diabetes -> Comorbidity (class) (Figure 28 and 31)

Recently, COVID-19 has severely affected diabetes patients creating respiratory problems and increased mortality rate [42]. Major features present in our datasets (Table 1) are outlined in Figures 4 and 5 showing the associative nature for the corresponding diagnosis whether Diabetes Mellitus (DM) or its comorbidities. Clinical notes are the readings taken from patients about symptoms felt related to health problems. These symptoms let the doctor think of possible causes of these symptoms and recommend tests to study the patient's condition. Results from these tests then help the clinician reach a possible diagnosis which is described to the patient in the form of practitioner comments. The cause of DM is the impairment to produce or respond to insulin that stabilizes the glucose level in the body. High level of glucose in the blood becomes the reason for forming different types of diabetes. The most common recommended tests for diagnosis of DM are HbA1c and glucose fasting (GF). Initial diagnosis may then be checked through some other tests to reach a final diagnosis and to predict complications that may give way to other related diseases as in the case of DM. Prescribed medicines may also result in the formation of certain skin or related allergies or diseases. This whole diagnostic process becomes input features for the deep learning neural networks that we chose in our experiments, that is, deep multinomial learning and LMHFL semi-supervised hybrid methods. The design and architecture of these proposed analytics are defined in section 6.


**Figure 5.** Deep Neural Networks Rule (A).

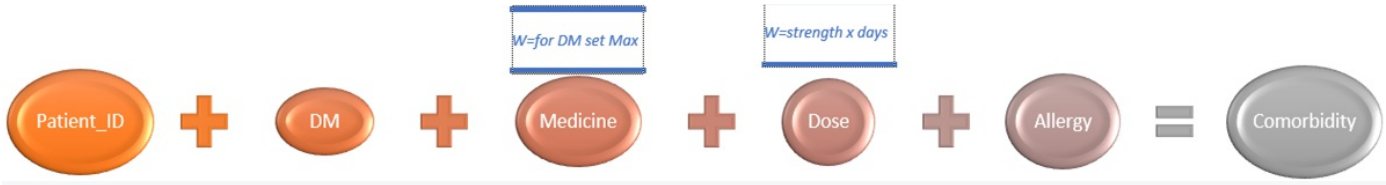



**Figure 6.** Deep Neural Networks Rule (B).

## 5. Datasets Pruned and Purified for Better Diagnostic Accuracy

Three different-sized datasets of 100 and 14407 endocrine patients' results were viewed for respective labels for 'Diagnosis'. Results were seen to have multiple labels for the same disease (Figure 1). Disease was either misspelled or was abbreviated or had different formats as 'hormonal', 'Harmonal' or 'Hormonal' and 'Hypertension' or 'HTN' or 'htn'. These impurities in datasets were corrected and each disease 'Diagnosed' was categorized into one label as classified in Table 2. Undiagnosed records were pruned and instances that remained were 2844, 9304 and 33185 respectively of three datasets that we analyzed in our experiments.

Null values or diagnoses labeled as 'For Labs', 'For Reports' or 'wrong entry' were categorized as 'Undiagnosed' and pruned to go through analytics as train and test data.

It is assumed that as accurate and pure the datasets, analysis results would be as fair and accurate.

Rule to classify all multiple labels for the same diagnosis into a single label in ICD-10 format, hence, setting multiple target values would be:

Diagnosis (labels misspelled or with different formats) -> Diagnosed (Unique Disease Labels) - ICD-10 codes

**Table 2.** Three endocrine datasets with the 'Diagnosed' label categorized for diagnosis of DM and other comorbidity diseases are listed with the frequency of occurrences.

| | **Dataset 1: 3650 instances of 100 patients Labeled as "Diagnosed"** | **Dataset 2: 15696 instances of 100 patients Labeled as "Diagnosed"** | **Dataset 3: 87803 instances of 14407 patients Labeled as "Diagnosed"** |
|---|---|---|---|
| **1** | 'UNDIAGNOSED': 806, | 'UNDIAGNOSED': 6392, | 'UNDIAGNOSED': 54618, |
| **2** | 'DM': 2359, | 'THYROID': 377, | 'Steroid Induced Diabetes': 1098, |
| **3** | 'THYROID': 248, | 'PRIMARY HYPOTHROIDISM': 18, | 'Gestational Diabetes' [6], [49]: 352, |
| **4** | 'HORMONAL': 198, | 'CV ASSESSMENT': 18, | 'DM': 18280, |
| **5** | 'THYROID DISORDER': 3, | 'HYPOTHYROIDISM': 36, | 'COPD': 1532, |
| **6** | 'GROWTH': 36 | 'INCREDIBLE STUDY': 36, | 'THYROID' [21]: 3384, |
| **7** | | 'MINIMAL CAD': 36, | 'SUBCLINICAL HYPOTHYROIDISM' [25]: 450, |
| **8** | | 'DM': 2690, | 'Prediabetes': 22, |
| **9** | | 'LBP': 36, | 'Primary Infertility': 10, |
| **10** | | 'HCC': 12, | 'Secondary Diabetes': 316, |
| **11** | | 'L4/5 DISC DEG': 4, | 'Primary Hyperparathyrodism' [25]: 285, |
| **12** | | 'HCC WITH METS': 2, | 'HORMONAL' [18]: 1698, |
| **13** | | 'CML': 2, | 'ANEMIA': 78, |
| **14** | | 'CHILD HOOD ASTHMA': 2, | 'Parkinson Disease Plus': 112, |

| | | | |
|---|---|---|---|
| 15 | | 'hearing problem': 43, | 'Toxic Epidermal Necrolysis secondary to Meloxicam': 124, |
| 16 | | 'TIA': 88, | 'Constipation': 12, |
| 17 | | 'BPH': 34, | 'LUNGS INFECTIONS': 54, |
| 18 | | 'LUTS': 54, | 'KIDNEY INFECTION': 72, |
| 19 | | 'acne': 26, | 'Urinary Tract Infection': 141, |
| 20 | | 'Sleep Starter (Jerk)': 22, | 'PNEUMONIA': 640, |
| 21 | | 'Paresthesias': 22, | 'Amiodarone associated Hyperthyroidism (Type 1)' [18]: 75, |
| 22 | | 'THORAT INFECTION': 22, | 'Pituitary Macroadenoma Treated with Transphenoidal surgery': 135, |
| 23 | | 'VERTIGO': 3, | 'Sheehan Syndrome': 276, |
| 24 | | 'Thyroiditis' [18]: 3, | 'Postural Drop': 114, |
| 25 | | 'HORMONAL': 299, | 'Vitamin D Deficiency': 60, |
| 26 | | 'HEADACHE': 3, | 'Stroke': 3038, |
| 27 | | 'PME ED': 12, | 'OSTEOARTHIRITIS': 120, |
| 28 | | 'ED': 12, | 'Lower Respiratory Tract Infection': 92, |
| 29 | | 'INFERTITILTY': 12, | 'Thyroiditis' [18]: 102, |
| 30 | | 'PME': 6, | 'Cellulitis': 10, |
| 31 | | 'CHOLECYSTECTOMY': 102, | 'Hypoglycaemia' [25]: 81, |
| 32 | | 'IRC': 423, | 'Obesity' [18]: 422 |
| 33 | | 'RHC pain': 34, | |
| 34 | | 'ANC': 219, | |
| 35 | | 'L lower parathyoid adenoma': 24, | |
| 36 | | 'Hypercalcemia, parathyroid adenoma,Thyroid Ca' [17], [50]: 24, | |
| 37 | | 'raised parathyroid gland': 24, | |
| 38 | | 'F/Up Thyroidectomy and Parathyroidectomy': 24, | |
| 39 | | 'Rt parathyroid adenoma excision': 66, | |
| 40 | | 'S/p Rt parathyroid excision.': 66, | |
| 41 | | 'Follow up case of Parathyroidectomy': 66, | |
| 42 | | 'parathyroid adenoma': 66, | |
| 43 | | 'HTN': 283, | |
| 44 | | 'SINUS TACHYCARDIA ? CAUSE': 69, | |
| 45 | | 'HEMATURIA': 16, | |
| 46 | | 'T PEDIS': 41, | |
| 47 | | 'GROWTH': 22, | |
| 48 | | 'Hyper lipidemia': 260, | |
| 49 | | 'CA BREAST': 2744, | |
| 50 | | 'BERY BERY': 49, | |
| 51 | | 'DYSPNEA': 49, | |
| 52 | | 'DRUG REACTION': 49, | |
| 53 | | 'FRECKLES': 201, | |

| | | | |
|---|---|---|---|
| **54** | | 'pruritus': 49, | |
| **55** | | 'sore throat-cough': 49, | |
| **56** | | 't pedis': 49, | |
| **57** | | 'S/P LAP CHOLE': 49, | |
| **58** | | 'NO OBVIOUS CV CAUSE IDENTIFIED': 49, | |
| **59** | | 'PSORIASIS': 54, | |
| **60** | | 'JOINT PAIN': 49, | |
| **61** | | 'BL RENAL STONE': 10, | |
| **62** | | 'thor': 10, | |
| **63** | | 'pain ear': 1, | |
| **64** | | 'ear wax': 1, | |
| **65** | | 'scabis': 5, | |
| **66** | | 'RAD/Allergic Rhinitis': 6, | |
| **67** | | 'GYNAE': 72 | |

Pruning was done manually on a need basis in parallel while conducting the experiments. On pruning the undiagnosed cases, big data set of 87,803 records was reduced to 33,185 instances. This big data set of 14407 endocrine patients was then divided into subsets separating each patient profile and disease profile. It was realized that each patient was diagnosed with one disease understood as a constant and would not be considered a target class whether it was either DM or a comorbidity disease. Where each disease profile had multiple patients.

Rule Sets from diagnosis of DM and its comorbidity made on these subsets were;

1. Patient_ID, Diagnosed, ICD-10-CM, Age, Gender, Note, PC, Examine -> Test
2. Patient_ID, Diagnosed, ICD-10-CM, Age, Gender, Note, PC, Examine, Test, Medicine, Strength, Days -> Allergy
3. Patient_ID, Diagnosed, ICD-10-CM, Age, Gender, Note, PC, Examine, Test, Allergy -> Medicine

## 6. Proposed Analytics Design and Architecture

The researchers started to leverage deep learning for the proposed semi-supervised analytics models. Our proposed analytics are semi-supervised as in Figure 2 we elaborated the design of these models through the hybridization of supervised and unsupervised ML algorithms. The datasets taken are also labeled with several unlabeled attributes that the analytics are self-trained. The models are designed by:

1. First, analyzing the problem domain for diagnosis of DM and its comorbidity diseases,
2. A diagnostic data model was designed that mapped all features to HL7 FHIR v4.0 standard with target labels set to diagnosed disease and its ICD-10-CM code for universal, uniform and generalized application.
3. The datasets were provided in parts; (i) the first dataset contained records of 100 DM patients that suffered from hormonal or thyroid problems as comorbidity, (ii) the second dataset was extended for the same patients with added

details of lab tests and comorbidity diseases, and (iii) finally, another dataset consisted of 14407 DM patients with some comorbidity diseases and added details of allergies and medication.

4. These datasets were transformed into three metadata sheets in Excel, were rigorously preprocessed in parallel, and were fed into analytics for better inferences and accuracy.
5. The custom rules were extracted and finally a uniform set of 9 features was taken for evaluation that had one target class Diagnosed and a Meta Attribute ICD-10-CM to label.
6. These deep learning heuristic models gave quite significant results with good accuracies if not definitive to avoid over-fitting.

### 6.1. Model 1: Architecture of LMHFL Analytics Heuristics Design

Researchers explored Louvain clustering for designing the big data analytics for the diagnosis of DM and its comorbidities. Louvain is known to detect communities in huge networks. This method has been used before to process 100 million nodes and links with success as it solves the NP-hard problem. Analysis of around 2 million nodes only takes 2 minutes on a PC. Louvain is efficient in finding communities, sub-communities and sub-sub-communities on zooming in. It is said to be a greedy optimization method to increase modularity in a network. Modularity is a measure to strongly group or cluster the communities in graph-like networks connected with edges. Networks of high modularity are densely connected but the nodes of the same cluster have sparse connections with nodes in other clusters. This optimized algorithm works in two steps. First, it looks for small communities, as in our case each diagnosis denotes a single community. In the second step, it connects all the nodes belonging to the same communities. This procedure is followed iteratively until maximum modularity is achieved forming a hierarchical structure. These graph-like communities are best to present associations within nodes as in our case diagnostic features make a single node that represents a diagnosis in the MDS graph (Figure 7).

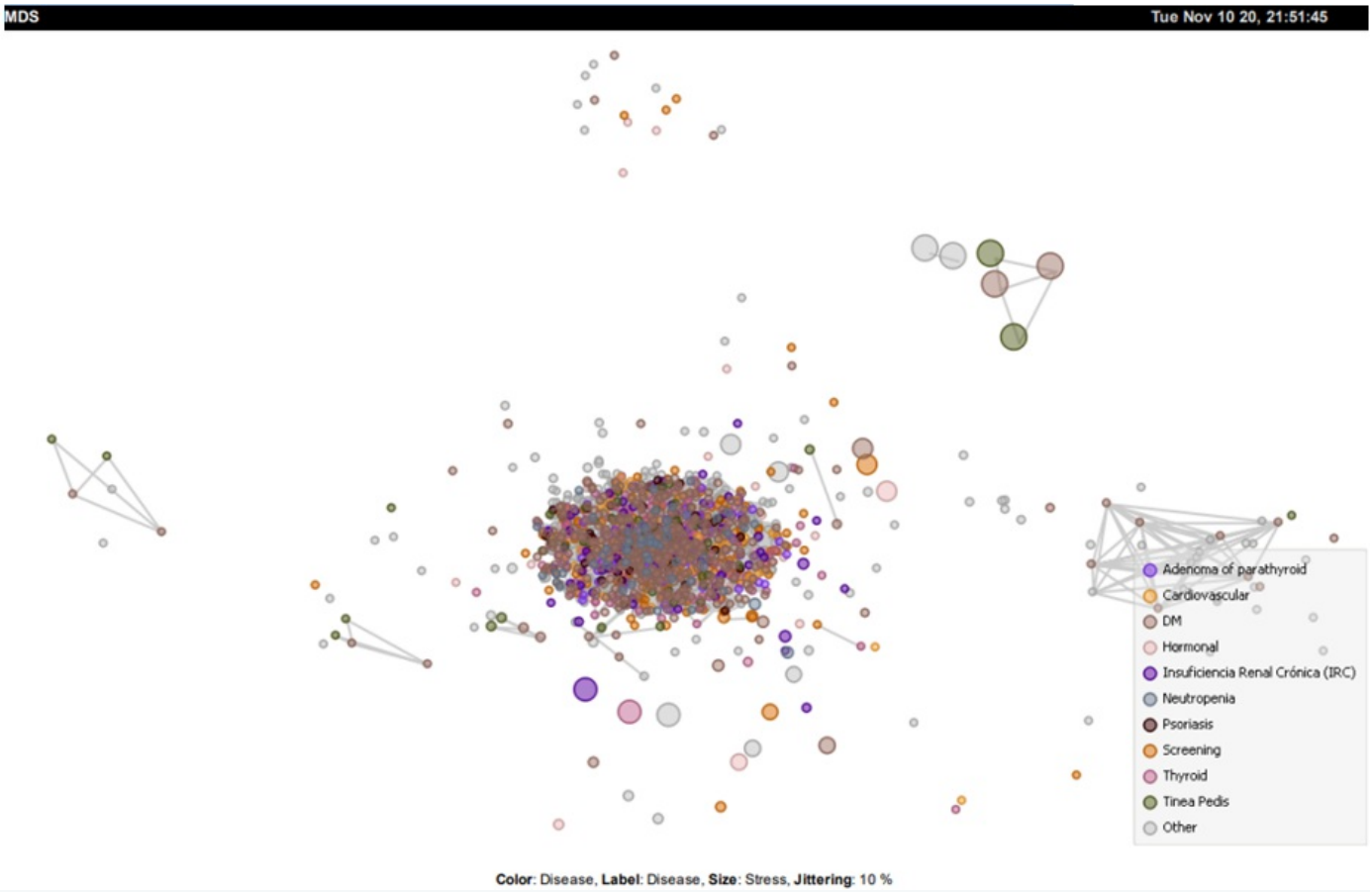



**Figure 7.** Visual graphical representation in Multidimensional Scale (MDS) from our proposed LMHFL model of a single patient diagnosed with DM and other diseases referred to as comorbidities.

A close connection is seen in multiple diagnoses for a single patient located on the graph based on a weighted average as per the importance and value of the diagnostic parameter or feature. The exact computational complexity is not known but the method runs in time O(nlogn) where most of the effort is put in the first step. Its strength lies in handling large data nodes in a very minimum time. The limitation is only due to the storage capacity as it is known that it successfully computed 118 million nodes in only 2.5 hours. Previously, its use was seen in studying social networks; Twitter and LinkedIn or audio sharing networks, mobile networks, YouTube, citation networks, Flickr or even human brain function networks, etc. [43]. Louvain was first published in 2008 by Blondel et al. for auto-detection of communities in large networks. Later in 2012, authors proposed some heuristic improvements to Louvain but unfortunately, they were constrained with time [44]. After this, there is some considerable work being seen to enhance it and fully connect networks [45].

1. Louvain Mani-Hierarchical Fold Learning

   In our work, we ensemble Louvain with Manifold and Hierarchical learning algorithms as semi-supervised for the best solution in a well-connected graph for rules induction and diagnostic accuracy evaluated by confusion matrix and F1-score and other measures like AUC, precision and recall, etc.
2. Fast-LMHFL Model

   Fastai emphasizes a lot on contributing to generalizing any model for uniformity as in our case we need it to fit with the

LMHFL model to optimize the analytics model run on our standardized big dataset built on HL7 FHIR v.4 and labeled with ICD-10-CM diagnostic codes for uniformity.

Fastai [46] a consistent deep learning API layered in four stacks was recently proposed that was worked upon from an idea that started to emerge in the year 2012 by Howard et al and his team of researchers at the University of San Francisco [47]. Fastai started with the launch of an easy-to-use use library with minimal code for deep learning and the course available for free at https://course.fast.ai/ and https://docs.fast.ai/. There is rapid research undertaken in recent years by fastai in the emergence of deep learning techniques for image and textual analysis using Named Entity Recognition (NER) concepts applied through spaCy known for Neural Linguistic Processing (NLP) for part of speech (POS) tagging. Fastai has contributed to several business processes like health as in here in particular.

Fast.ai library is easily embedded in the Orange framework installed using the “conda install -c fastchan fastai anaconda” command. Fastai in comparison to other known NLP environments is easy to use and integrate with cloud platforms as with Orange in this paper. It promises to assist in the custom development of deep learning models on top of it without starting from scratch. This research, therefore, employed deep learning rich Fast.ai library and integrated with our LMHFL model to fine-tune and embed textual analysis to fields in our tabular dataset of DM patients suffering from other comorbidity diseases as identified earlier in the paper.

## 6.2. Model 2: Architectural Design of Deep Multinomial Distribution Model

The paper [39] focused on replacing the Bernoulli distribution with a multinomial as it found multinomial distribution is an efficient method to generalize the binomial distribution for success and failure by expanding it to some multiple graded scales from low to extreme scenarios. Sparse Bayesian Extreme Learning Machine (SBELM) had 3 layered neural networks with probabilistic models using Bernoulli distribution for binary classification and went on to do multi-label classification through pairwise coupling by binary classifiers. Uncertainty constraint was of concern; therefore, multinomial distribution was employed for unambiguous multiclass distribution and Multinomial Bayesian Extreme Learning Machine (MBELM) replaced Bernoulli with Multinomial distribution. SBELM used sigmoid activation which is designed for binary classification only. Sigmoid activation was replaced by softmax activation. It was proposed with the purpose of its application in real-time systems and therefore this paper employed a deep multinomial distribution model for experimentation.

Deep Multinomial Distribution employed for our heuristics for hybrid analytics has two versions; Deep Multinomial Distribution and Deep Multinomial Multi-label Learning Distribution.

1) Deep Multinomial Distribution

The architecture given in Figure 8 worked on H2O 3.30.0.1 as deep multi-layer feed-forward artificial neural networks having stochastic gradient descent with backpropagation.

In our setting it was trained on 10 epochs, activation function was tanhwithdropout with one input layer, two hidden layers

of sizes [50,50] or [32,35] and an output layer that gave the best results in terms of accuracy. Regularization measures L1 and L2 were set to 1e-5 and 0 respectively. The L1 regularization method constrained the absolute values of the weights and the net effect of dropping some weights (setting them to zero) from a model to reduce complexity and avoid over-fitting. The L2 regularization method constrained the sum of the squared weights. This method was to introduce bias into parameter estimates, but frequently produced substantial gains in modeling as estimate variance is reduced.

The adaptive learning rate was set as epsilon = 1e-8. The hidden dropout ratio was 0.5 and missing values were handled using mean imputation.

10 folds cross-validation through decision tree was chosen to train an example set of 33,185 rows of Dataset 3 (Table 1) with 18 columns; PATIENT_ID, VAN, VISIT_DATE, AGE, GENDER, EXAMINE, TEST, RESULT, ALLERGY, NOTE, ASSESSMENT, PC, ICD-CODE, MEDICINE, STRENGTH, UNIT_DESCRIPTION, DAYS, Diagnosed.

Stratified sampling was applied in parallel execution for batch processing.

100% accuracy was predicted using a confusion matrix having a minimum log loss of 0.0843 in Figure 10.

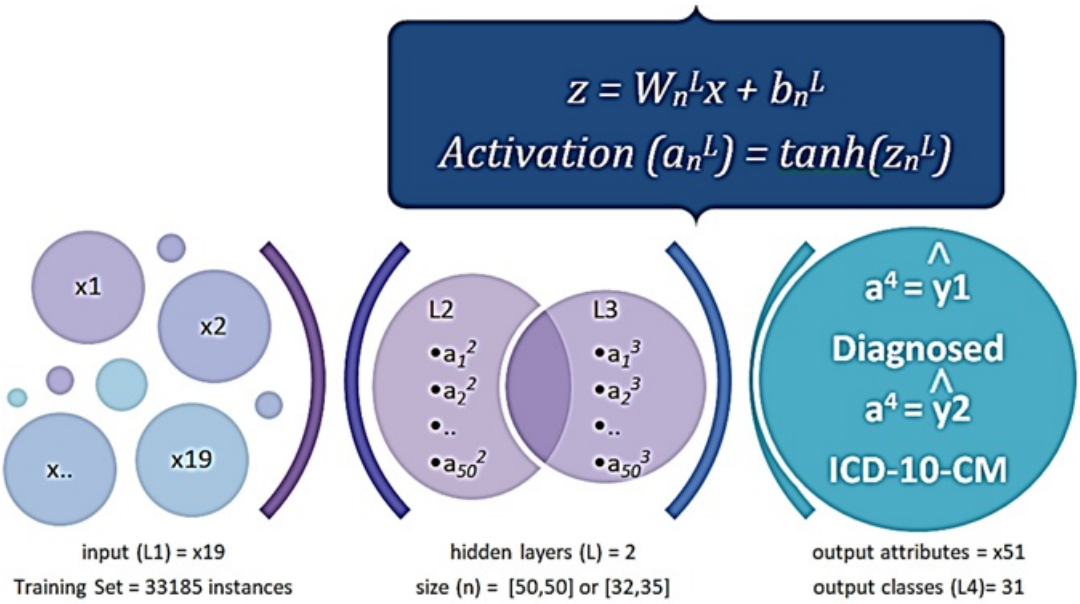



**Figure 8.** DNN diagram showing the architecture of our heuristics approach for analytics.

2) Deep Multinomial Multi-label Learning Distribution

The results became better with log loss accuracy to 0.0711 (Figure 10) when deep multinomial distribution was hybridized with the multi-label model.

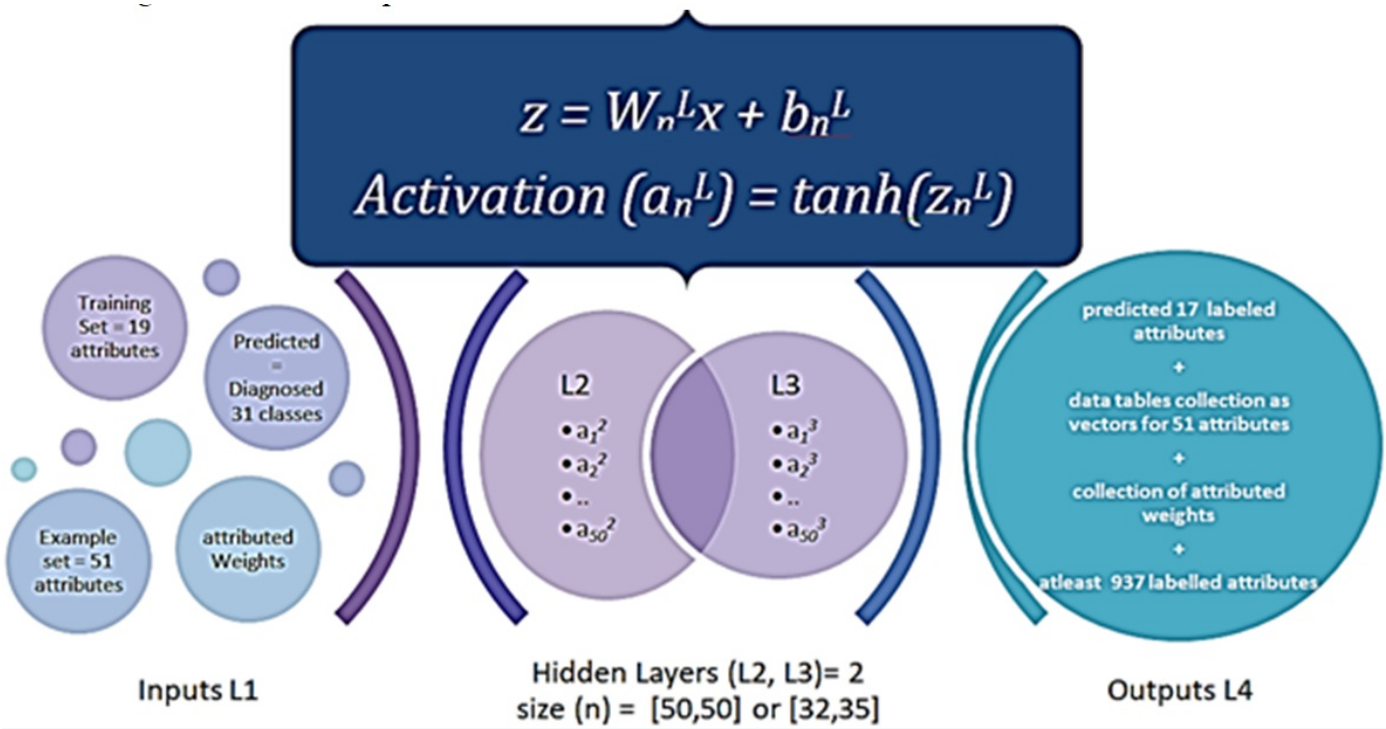



**Figure 9.** Deep Multinomial Multi-label Learning Model Architecture with 4 layers showing input training set attributes and output labelled Diagnosed class attributes.


**Figure 10.** Deep Multinomial Model validated by cross-validation is input into Multi-label Model and validated again using a decision tree.

Deep Multinomial Model used two labels ‘Diagnosed’ and ‘ICD-10’ to predict the ‘Diagnosed’ class. The prediction model is cross-validated using a decision tree and inputted into the Multi-Label Model. In the Multi-Label Model, we filtered all attributes leaving behind Patient-ID (as it is an identifier only and was used as a batch attribute in the prediction model).

The 17 labeled attributes (Figure 11) now train the sub-process using split validation to predict each individual label attribute and predicted vector dataset as an output with at least 937 weighted attributes.

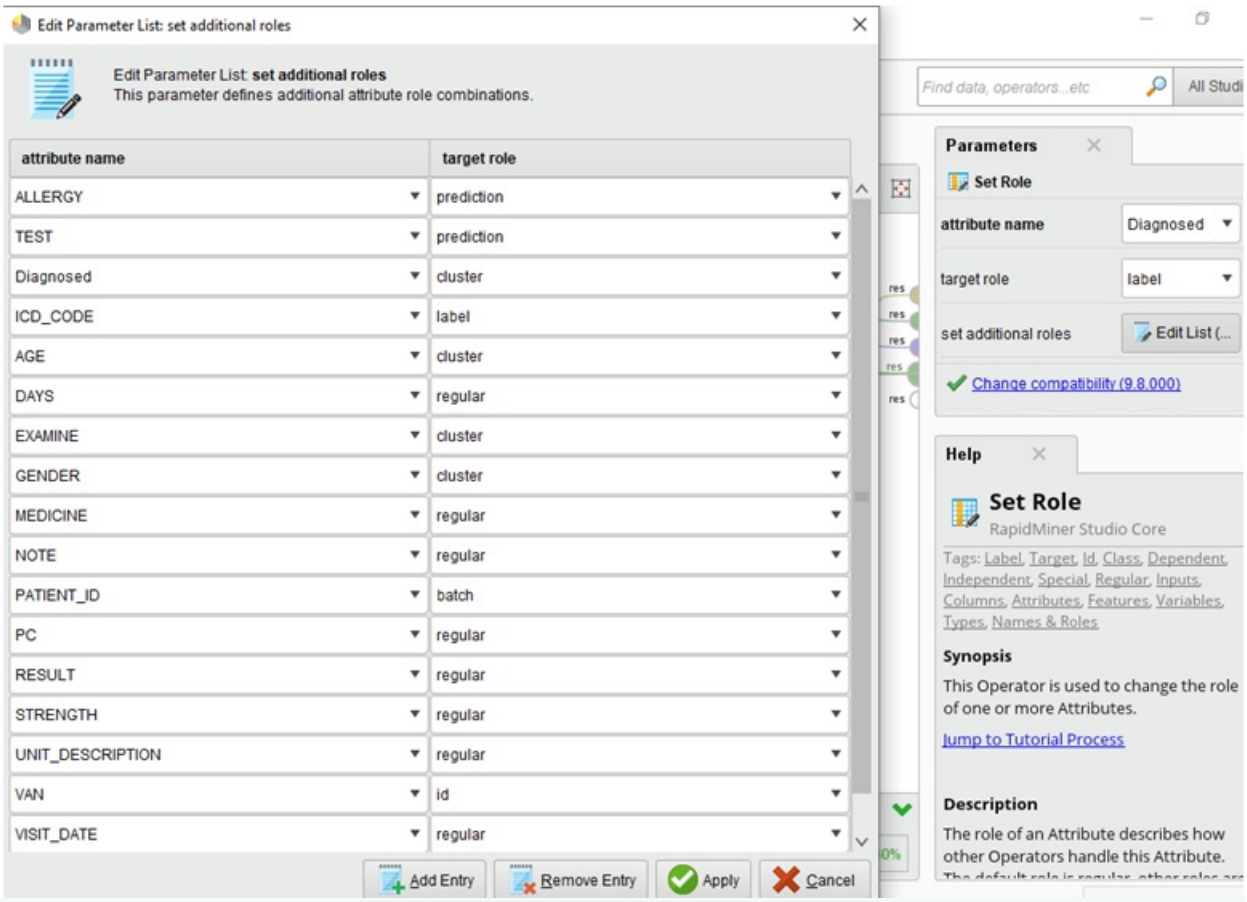



**Figure 11.** 17 Labeled attributes taken in training set with Patient-ID as batch identifier to predict multiple labelled weighted attributes for prediction of Diagnosed and ICD-10 code classes.

## 7. Experiments

This paper experimented with endocrine datasets on Orange[48], Qlik and RapidMinor Commercial trial versions as well as educational/student software designed specifically for researchers., These platforms have several advanced analytical ML algorithms and tools available to experiment and evaluate in terms of speed and accuracy. (i) Orange is a data mining component-based framework for visualization and analysis found as part of the Anaconda platform that may be used by a novice or an expert, (ii) Qlik Sense is also an open-source cloud platform freely available to read and visualize data in statistical form, and (iii) Rapid Miner free edition also comes with several extensions for data scientists to build and optimize machine learning models with automation. We proposed two analytical models; LMHFL and Deep Multinomial Distribution. Each of these models has a modified version that optimizes its performance in terms of accuracy.

Around 245 results in visual form as .pdf documents or images were compiled to evaluate the analytics models discussed in section 5. Two endocrine datasets were taken of 100 and 14407 patients with multiple visits to study the risks formed and occurrences of comorbidity diseases relative to DM. This data was transformed into metadata applying the data

model based on HL7 FHIR v4 and ICD-10-CM coding scheme used specifically for diagnostics in the United States with authorization of WHO.

### 7.1. Case 1: Model 1. LMHFL Analytical Model Applied in Orange Framework

Orange 3 was used for data visualization through analysis using Python at the backend. Various analytical algorithms for clustering, classification and association were tried and many data views were generated to show results for 3650 instances from multiple visits of 100 endocrine patients [48].

The LMHFL model, and Fast-LMHFL; its modified version integrated with fast.ai for text analysis with maximum efficiency built was Manifold Hierarchical Cluster (MHC) combined with Louvain clustering algorithm [49]. The efficiency is evaluated on observation derived from accuracy measures (Table 3). The resultant predictions and accuracies (Table 3) are explained in section 8 and later. Many inferences and co-relations were found with multiple patient profiles on 100 patients datasets as in Figures 12 to 15.

From several results here few are shown in Figures 12, 13, 14 and 15. The dataset with a limit on the size of instances gave us various patient-specific views in relation to features extracted of common comorbidities formed due to DM as ‘hormonal’ and ‘thyroid’ diseases in dataset 1 (Table 2).

Figure 11 clearly relates patients’ age groups and recommended exams with target results to the respective diagnosed disease. It is seen that DM patients are found mostly in the ages of 40 to 60 and later in the 70s. Through Principal Component Analysis (PCA), in Figure 15, it is seen that age and gender factor plays an important role in isolating DM patients from the rest as females with obesity are more affected in later ages forming further complications with time and negligence or left undiagnosed.

The sensitivity of the analytical model on Orange is seen, in Figure 12, to 14, for its identification of multiple similar labeled clusters as ‘Harmonal’ and ‘Hormonal’ or ‘Thyroid’ and ‘Thyroid Disorder’ due to errors in typing during the assessment of the patient. It also separates wrongly labeled patients as ‘Undiagnosed’ with incomplete diagnoses as ‘blank’, ‘Growth’ or ‘For Lab’. A detailed description of Diagnosed classes can be seen in Table 2 section 5.

A patient-specific detailed profile is highlighted categorically through colors for disease labels and symbols for gender in Figure 12 clustered on a two-dimensional scale. Views are seen clearly separated by parameters on scatter plot within a two-dimensional graph where the x and y axis may be altered for different views of the similar diagnosis (Figure 12, 13, 14 and 15).


**Figure 12.** Exam Results of Male/Female Patients with age showing mostly Female population diagnosed with Diabetes Mellitus (DM) between ages of the 40s-60s and after 70s. Males are forming DM in 50s and late 60s which is very less.

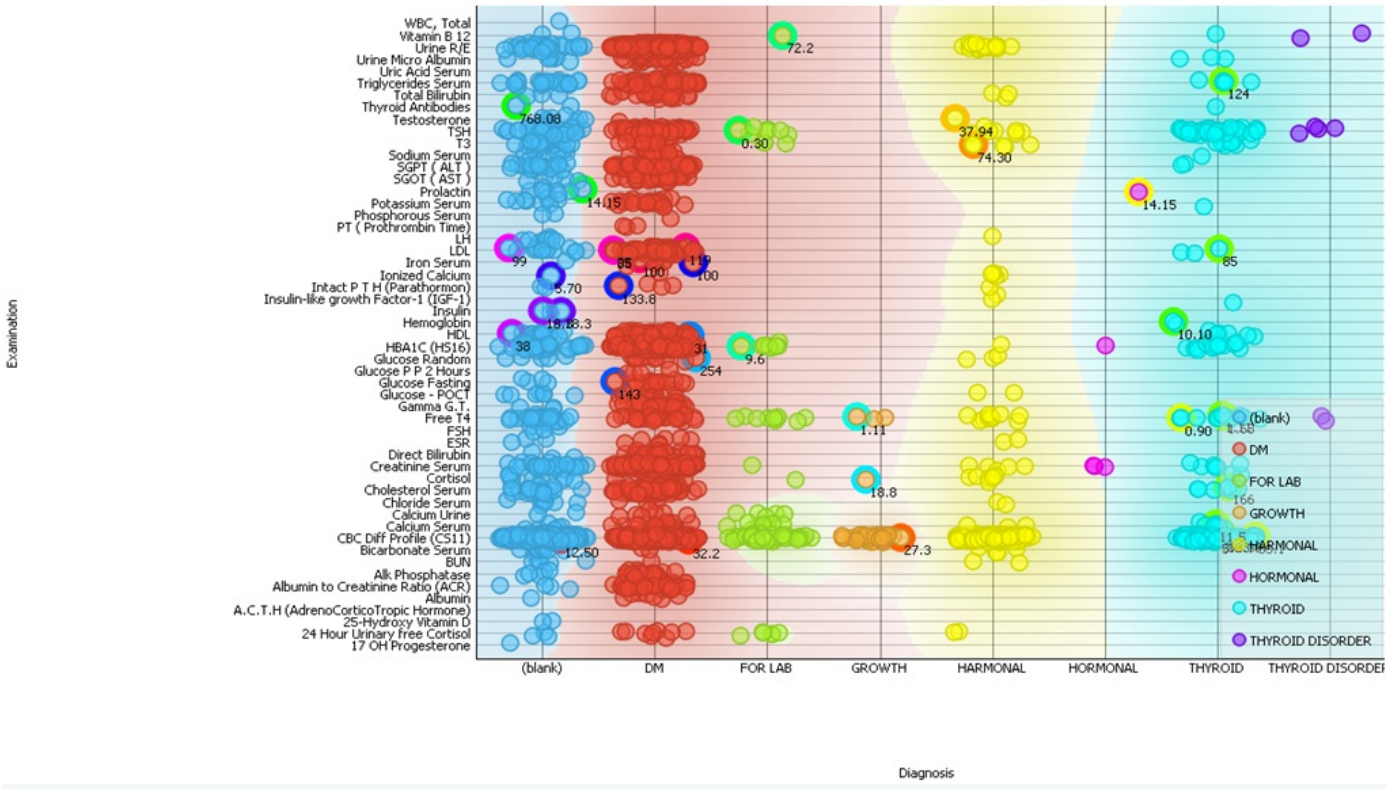



**Figure 13.** Test Results of Diabetic Patients for different diagnoses (some clusters are mislabeled or given similar labels like ‘For Lab’, ‘Growth’ or Hormonal due to human errors).

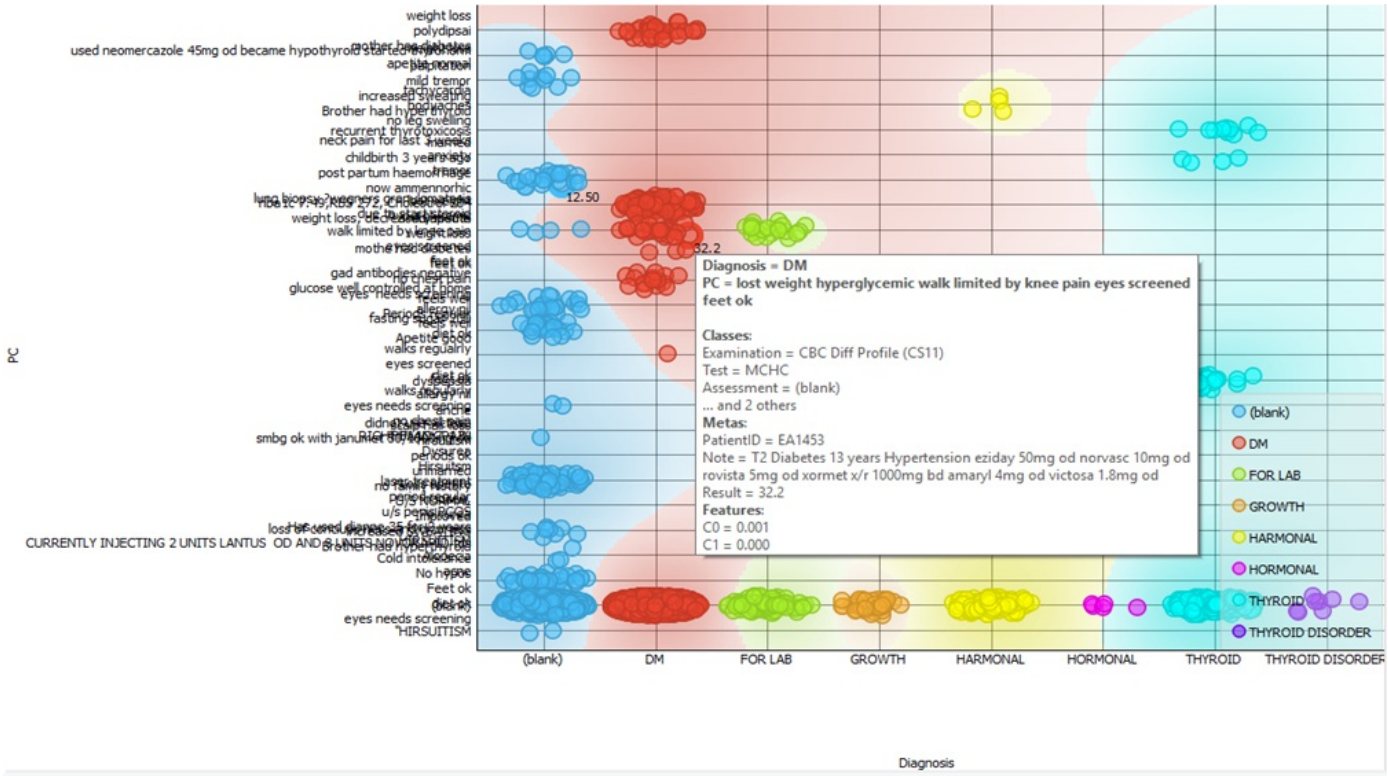



**Figure 14.** Patient's profile with DM displaying Practitioner's Comments with MCHC test=32.2 for a given Exam "Complete Blood Count (CBC)".

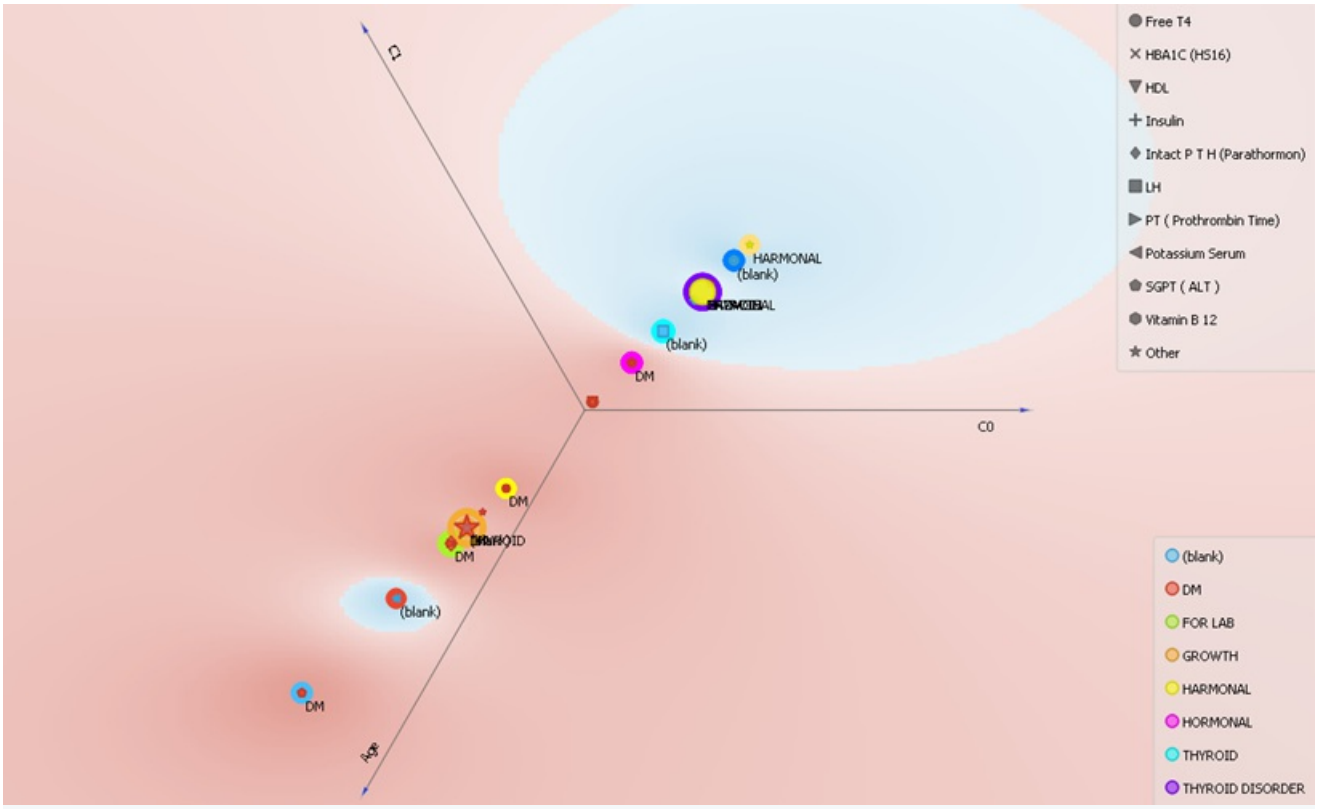



**Figure 15.** Component Analysis using PCA. Here again, it is seen that DM is found in later ages.

A relatively larger dataset of 15696 instances of 100 endocrine patients with multiple diagnoses on multiple visits gave a limited view that was not patient-specific. Researchers used this dataset to see a broader pictorial relational view of

chosen features to judge the performance and accuracy of Louvain Mani-Hierarchical Fold Learning (LMHFL) [48]. It was tried with different model settings on individual patient profiles and combined with varying hyperparameters and showed precise hierarchical clusters, scatter plots and multidimensional views to distinguish between DM and its comorbidity diseases. The accuracy and precision shown in Table 3 highlight the correlations between features, reaching up to 0.989. Additionally, rules extracted using the Laplace achieved a maximum accuracy of 0.7, whereas the Entropy measure failed. Processing time to run analytics and get results using the LMHFL model in Orange framework for 9646 records of 100 patients spanned over 48 hours. Therefore, we performed analytics on individual patients' profiles (Section 10).

### 7.2. Case 2: Statistical Analytics Applied in Qlik

In Qlik, a normalized dataset of 100 diabetic patients was generated through multiple tables (Figure 16); Patients, Tests, Diagnosis, Medication Statement and Allergies.

The endocrine diseases that did not show in Orange Framework due to few records are seen in QLik in Figures 18 and 19. Therefore, it is understood that Orange Framework is designed for relatively bigger data than QLik having ML capabilities, whereas Qlik is specifically for statistical views to build an understanding of the data at hand. This dataset generated views based on different statistical analysis methods. Bar chart views were created to separate endocrine patients based on gender, age, and mean age of a diagnosed patient in Figures 17 and 18. Figures 18 to 22 show several forms of DM seen as diabetes, diabetes mellitus or type 2 diabetes. In Figure 18, DM is seen affecting at ages of 19, 30s, 40s, 60s and above 70. T2D is found at the age of 60. Figure 19 shows all diagnoses records against the total counts for each. Then there is a scatter plot for Dataset 3 of 14407 patients in Figure 20 showing larger population is diagnosed with DM around the age of 50 with some other comorbidity diseases. Figure 21 shows the gender stats for patients diagnosed with DM that are almost equal for males and females both. In cases diagnosed with comorbidity diseases there are differences in these ratios as for hormonal, obesity, kidney infection, hypoglycemia and Parkinson's disease female respondents are seen. Male patients are seen to have lung disease i.e., Chronic Obstructive Pulmonary Disease (COPD). In Figure 22 combined effect of gender and age is seen on the diagnosed diseases. Tree Maps or Heat Maps gave a relational limited view of patients' data in Figures 23 and 24, grouping the patients with respective diseases based on test results and clinical notes.

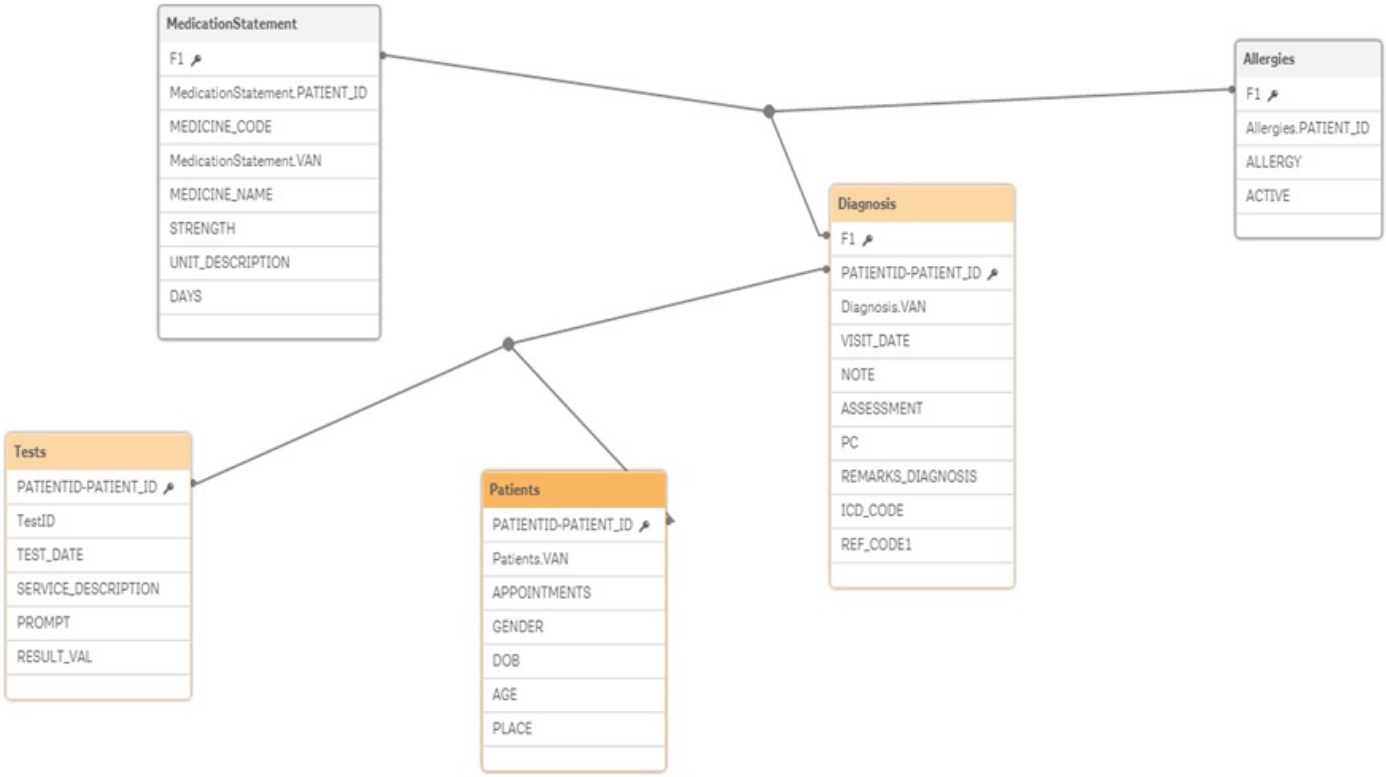



**Figure 16.** Normalized Dataset of 100 diabetic patients having multiple visits records.

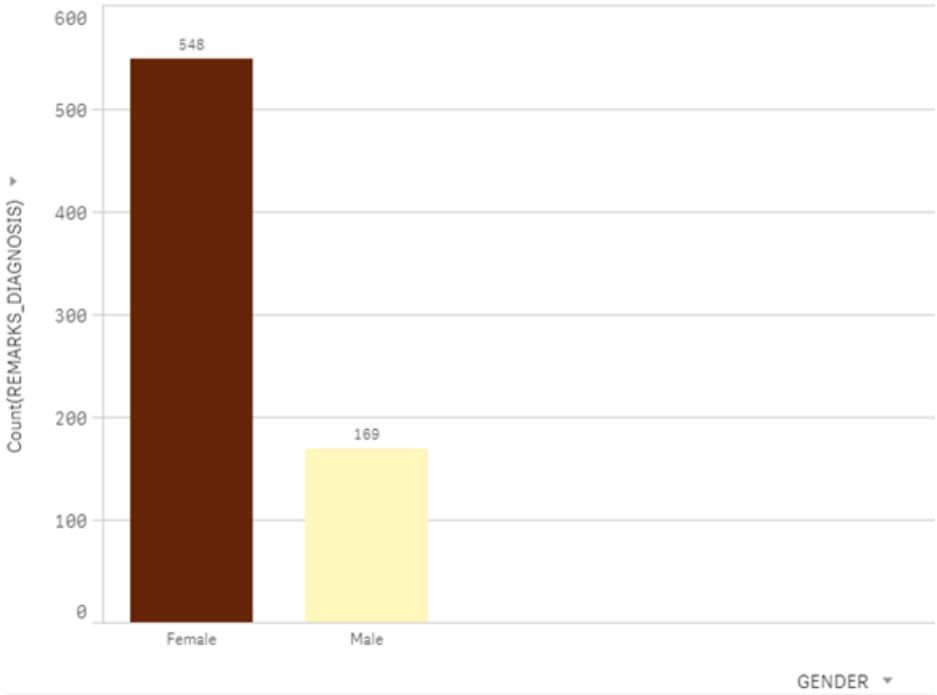



**Figure 17.** Clear Statistical view of diabetic females diagnosed more compared to men.

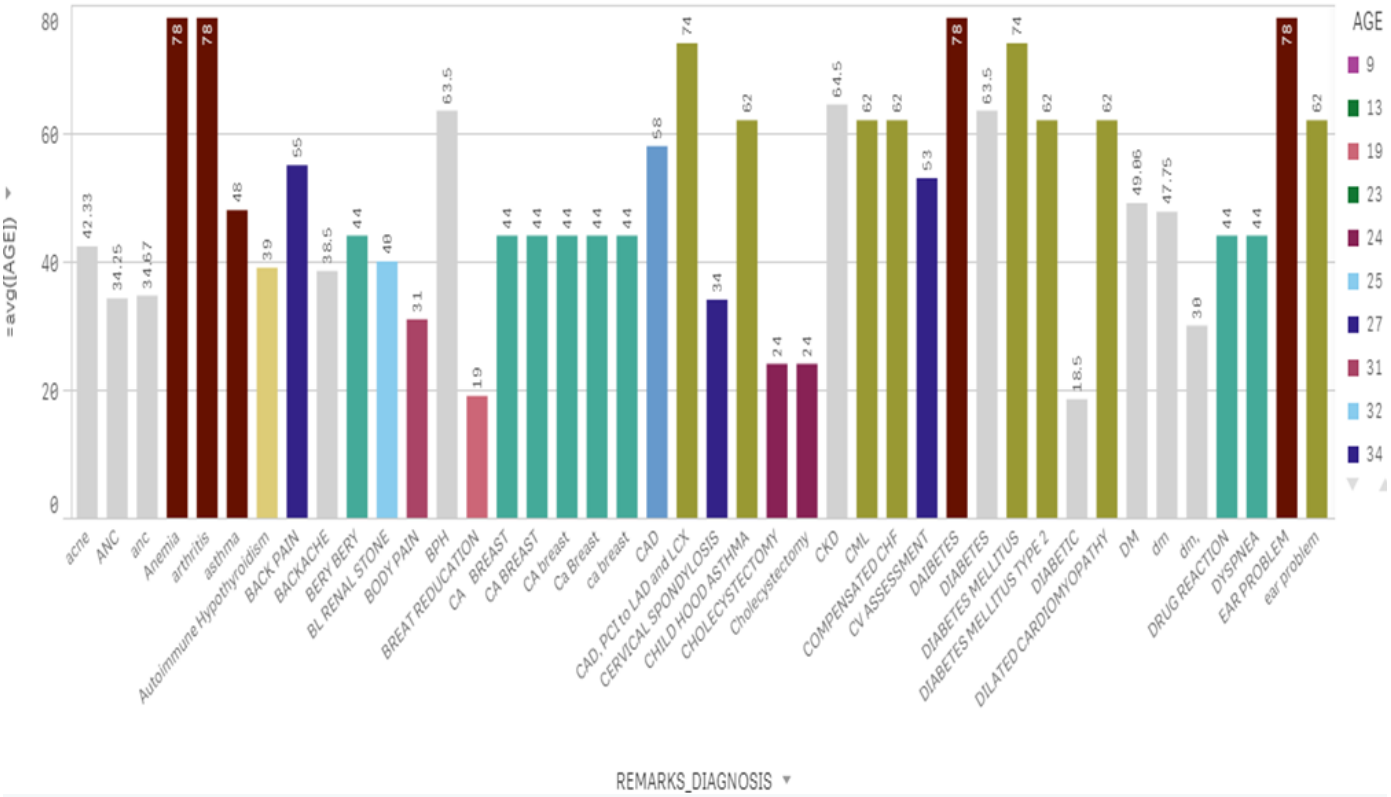



**Figure 18.** Statistical view showing mean ages for different forms of diagnosis in diabetic patients.

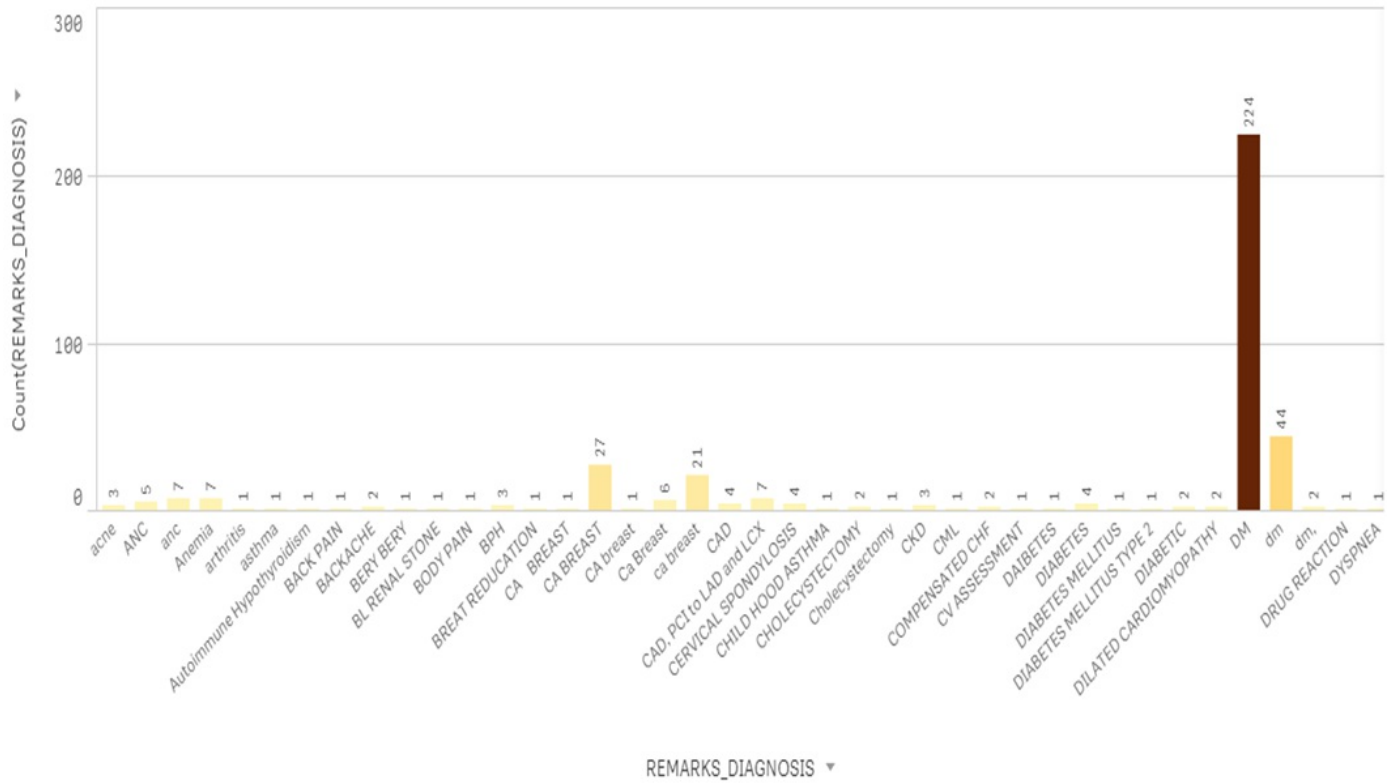



**Figure 19.** Patients with DM are more compared to other types of diabetes in given dataset 2 of 100 patients.

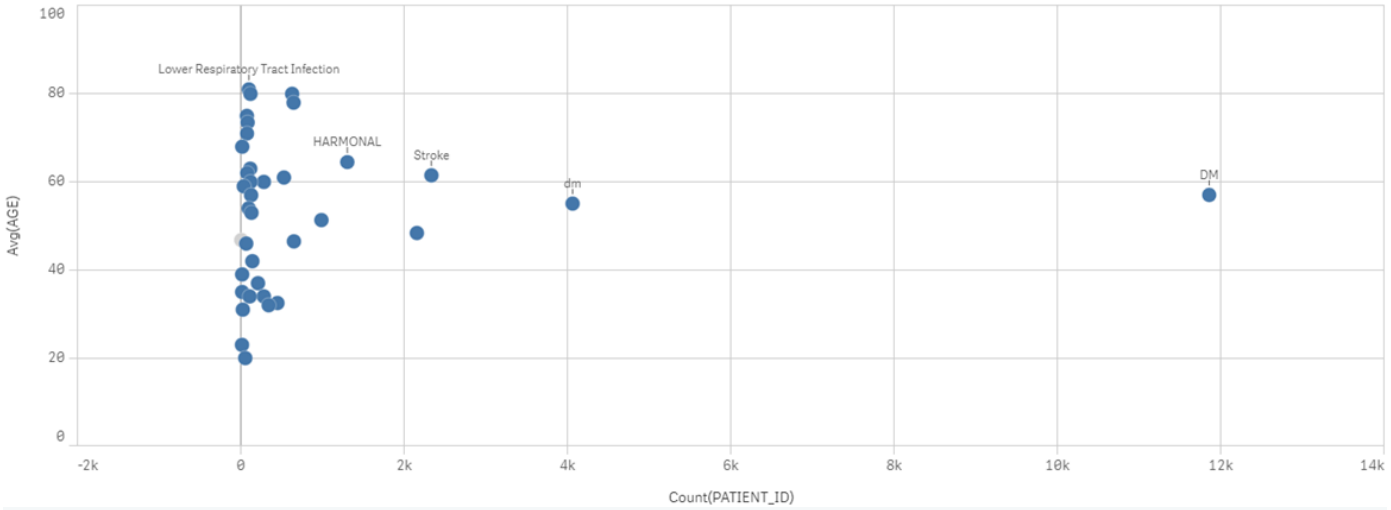



**Figure 20.** Scatter plot of 14407 Patients showing larger population diagnosed with DM having a mean age of 50 plus and stroke and other underlying diseases occurring after the age of 60.

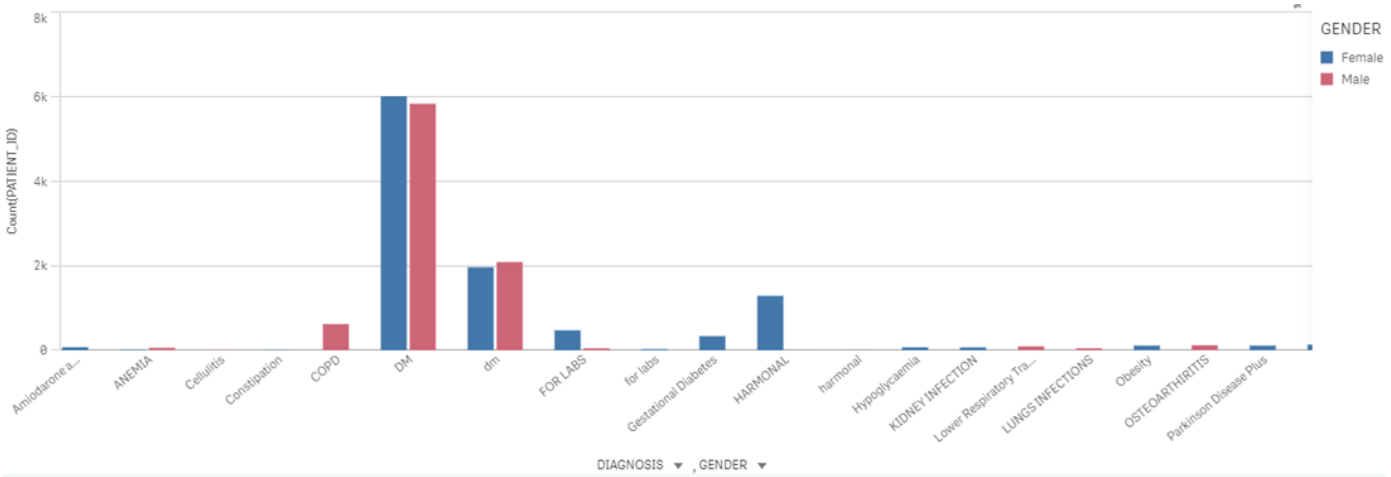



**Figure 21.** Dataset 3 of 14407 patients have equal ratio of male/female DM patients.

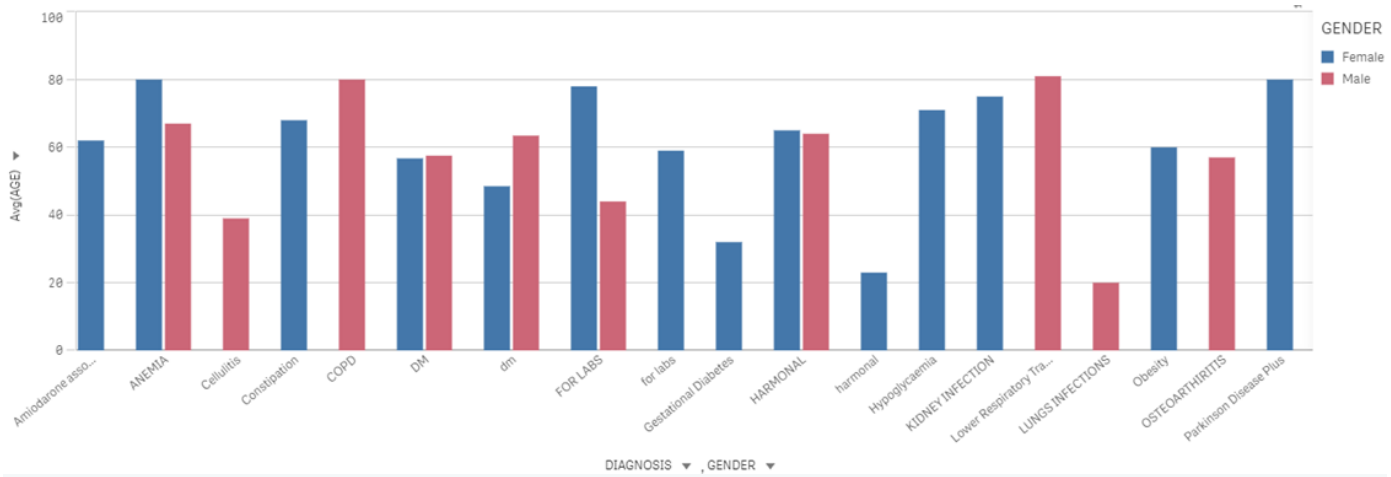



**Figure 22.** 14407 endocrine patients distributed on a scale of gender and age.


**Figure 23.** Graphical representation of Tree for population diagnosed with DM for results formed for various tests predicted based on clinicians notes (as for Lymphocytes results lies in 23 to 28, LDL is 174, Hemoglobin ranges from 11.3 to11.9 mostly and HbA1c lies between 8.1 to 9.1 for DM).

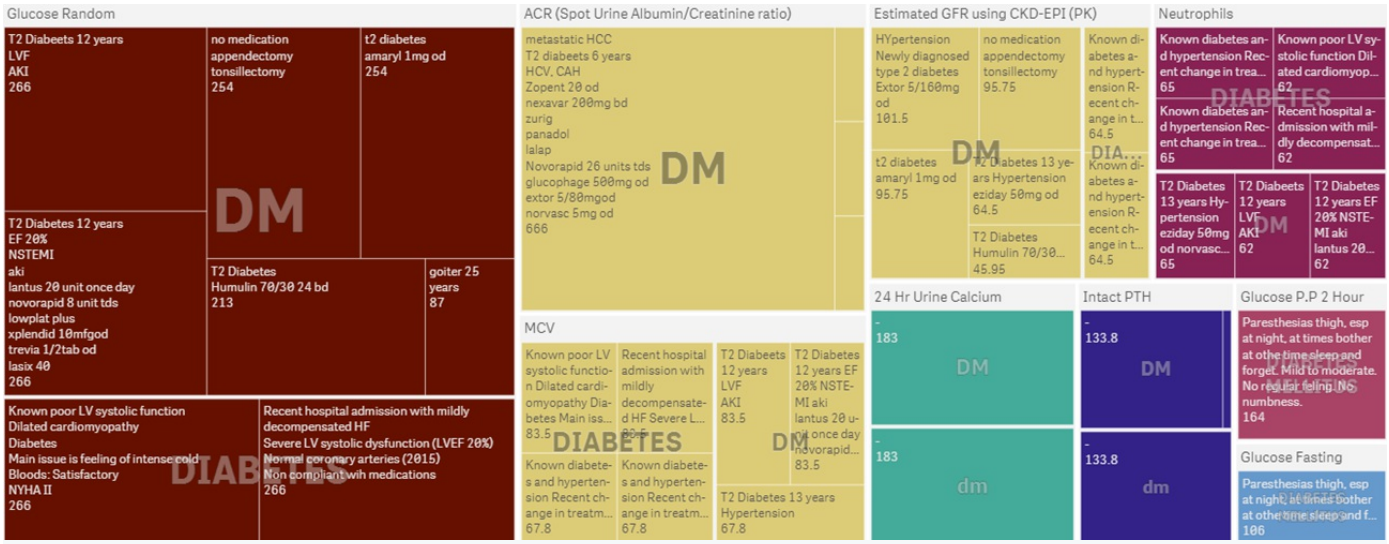
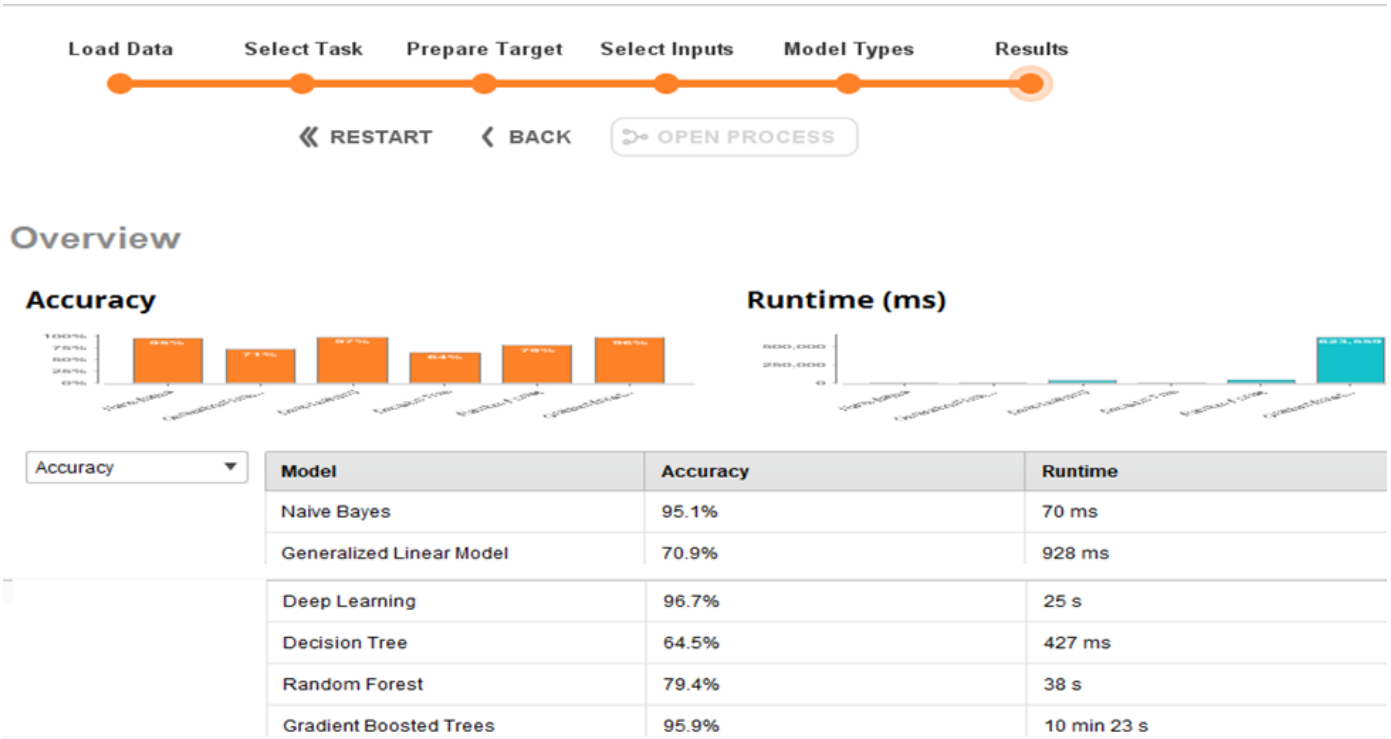



**Figure 24.** Graphical representation of Tree for population diagnosed with DM for results formed for various tests predicted based on clinicians notes (as for Glucose Random with a range of results from 87 to 266, GFR lies in 45.95 to 101.5, 24-hour Urine Calcium is 183 for DM patient).

## 7.3. Case 3: Model 2. Deep Multinomial Distribution Analytics Applied in RapidMiner Framework

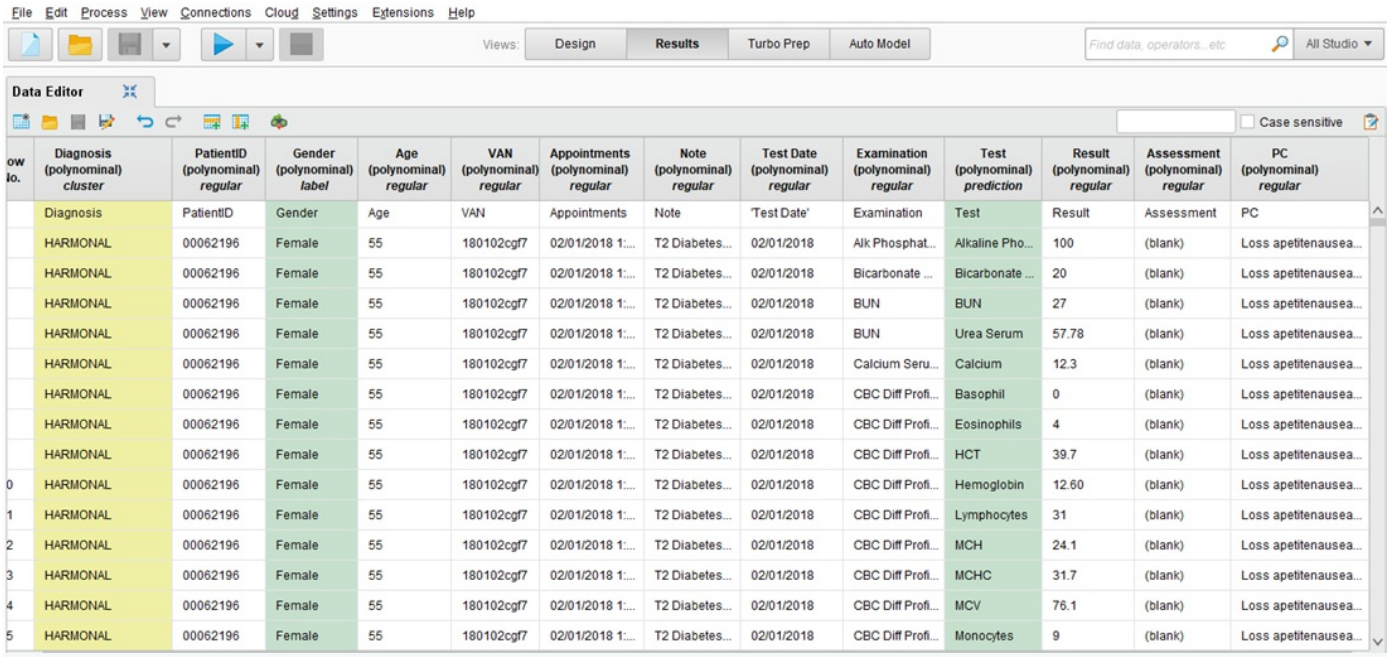
RapidMiner Commercial application with multiple ML analytical algorithms was done on 100 patients Dataset 1 with a size limited to 3650 instances. Two rules established were; Rule 1: PatientID, Gender, Age, VAN, Appointment, Note, Test Date, Examination, Test, Result, Assessment, PC -> Diagnosis (Labelled) (Figure 25) and Rule 2 (Figure 26 and 27): a) PatientID, Gender (Label), Age, VAN, Appointment, Note, Test Date, Examination -> Test (Prediction) b) Test (Predicted), Result, Assessment, PC -> Diagnosis (Cluster) Commercial RapidMiner Software gave 100% accuracy on 3650 records having multiple labels in Rule 2 as ‘Test’ prediction and ‘Diagnosis’ classification for analytics using Naïve Bayes in 87ms, Random Forest in 14s, deep learning with runtime of 23s and Gradient Boosted Trees (GBT) taking 2mins 44s processing

time (Figure 27).



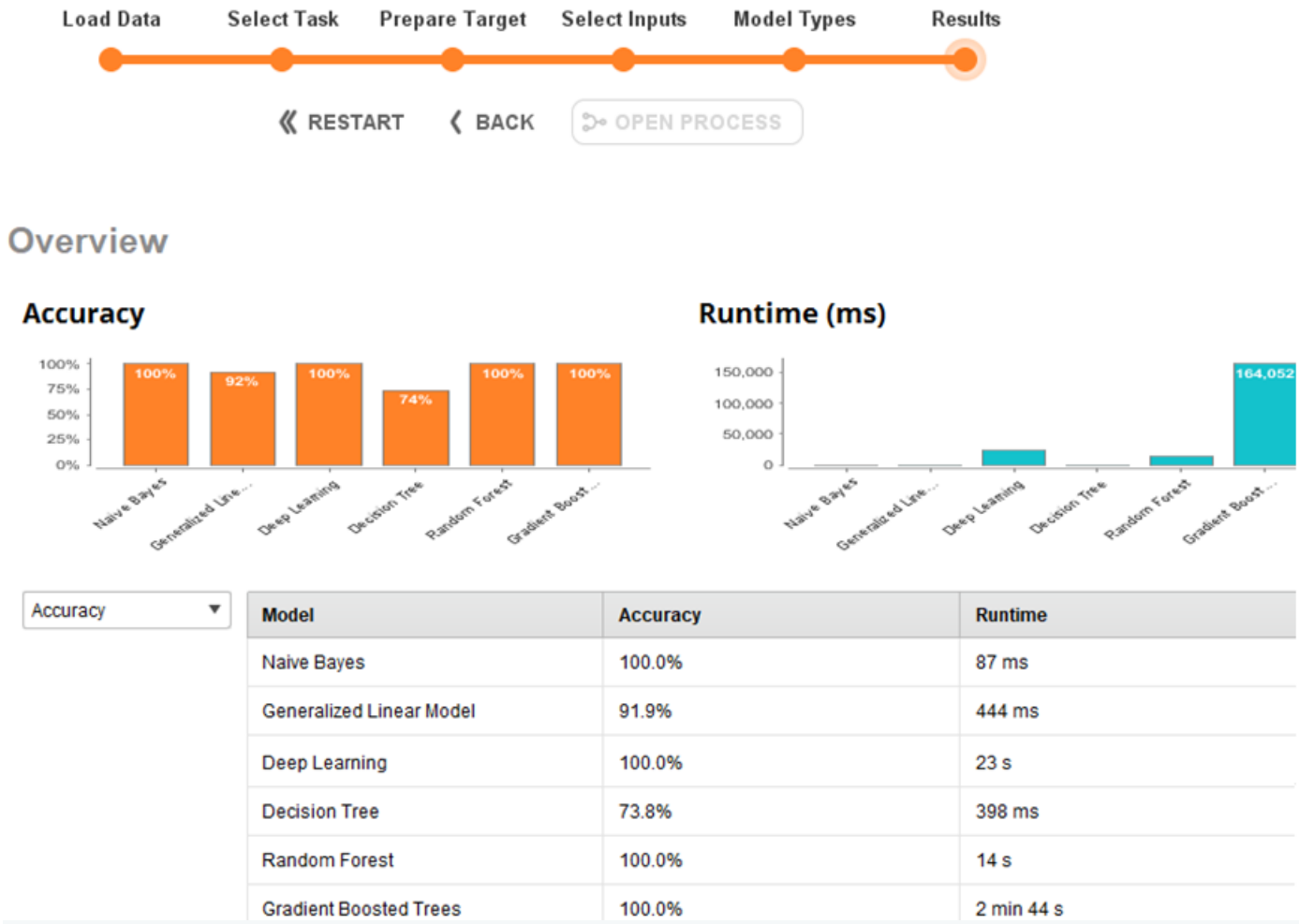

| Model | Accuracy | Runtime |
|---|---|---|
| Naive Bayes | 95.1% | 70 ms |
| Generalized Linear Model | 70.9% | 928 ms |
| Deep Learning | 96.7% | 25 s |
| Decision Tree | 64.5% | 427 ms |
| Random Forest | 79.4% | 38 s |
| Gradient Boosted Trees | 95.9% | 10 min 23 s |

**Figure 25.** Performance of Classification results for Rule 1 on 3650 instances by applying known algorithms.




| ow lo. | Diagnosis (polynominal) *cluster* | PatientID (polynominal) *regular* | Gender (polynominal) *label* | Age (polynominal) *regular* | VAN (polynominal) *regular* | Appointments (polynominal) *regular* | Note (polynominal) *regular* | Test Date (polynominal) *regular* | Examination (polynominal) *regular* | Test (polynominal) *prediction* | Result (polynominal) *regular* | Assessment (polynominal) *regular* | PC (polynominal) *regular* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Diagnosis | PatientID | Gender | Age | VAN | Appointments | Note | 'Test Date' | Examination | Test | Result | Assessment | PC |
| | HARMONAL | 00062196 | Female | 55 | 180102cgf7 | 02/01/2018 1:... | T2 Diabetes... | 02/01/2018 | Alk Phosphat... | Alkaline Pho... | 100 | (blank) | Loss apetitenausea... |
| | HARMONAL | 00062196 | Female | 55 | 180102cgf7 | 02/01/2018 1:... | T2 Diabetes... | 02/01/2018 | Bicarbonate ... | Bicarbonate ... | 20 | (blank) | Loss apetitenausea... |
| | HARMONAL | 00062196 | Female | 55 | 180102cgf7 | 02/01/2018 1:... | T2 Diabetes... | 02/01/2018 | BUN | BUN | 27 | (blank) | Loss apetitenausea... |
| | HARMONAL | 00062196 | Female | 55 | 180102cgf7 | 02/01/2018 1:... | T2 Diabetes... | 02/01/2018 | BUN | Urea Serum | 57.78 | (blank) | Loss apetitenausea... |
| | HARMONAL | 00062196 | Female | 55 | 180102cgf7 | 02/01/2018 1:... | T2 Diabetes... | 02/01/2018 | Calcium Seru... | Calcium | 12.3 | (blank) | Loss apetitenausea... |
| | HARMONAL | 00062196 | Female | 55 | 180102cgf7 | 02/01/2018 1:... | T2 Diabetes... | 02/01/2018 | CBC Diff Profi... | Basophil | 0 | (blank) | Loss apetitenausea... |
| | HARMONAL | 00062196 | Female | 55 | 180102cgf7 | 02/01/2018 1:... | T2 Diabetes... | 02/01/2018 | CBC Diff Profi... | Eosinophils | 4 | (blank) | Loss apetitenausea... |
| | HARMONAL | 00062196 | Female | 55 | 180102cgf7 | 02/01/2018 1:... | T2 Diabetes... | 02/01/2018 | CBC Diff Profi... | HCT | 39.7 | (blank) | Loss apetitenausea... |
| 0 | HARMONAL | 00062196 | Female | 55 | 180102cgf7 | 02/01/2018 1:... | T2 Diabetes... | 02/01/2018 | CBC Diff Profi... | Hemoglobin | 12.60 | (blank) | Loss apetitenausea... |
| 1 | HARMONAL | 00062196 | Female | 55 | 180102cgf7 | 02/01/2018 1:... | T2 Diabetes... | 02/01/2018 | CBC Diff Profi... | Lymphocytes | 31 | (blank) | Loss apetitenausea... |
| 2 | HARMONAL | 00062196 | Female | 55 | 180102cgf7 | 02/01/2018 1:... | T2 Diabetes... | 02/01/2018 | CBC Diff Profi... | MCH | 24.1 | (blank) | Loss apetitenausea... |
| 3 | HARMONAL | 00062196 | Female | 55 | 180102cgf7 | 02/01/2018 1:... | T2 Diabetes... | 02/01/2018 | CBC Diff Profi... | MCHC | 31.7 | (blank) | Loss apetitenausea... |
| 4 | HARMONAL | 00062196 | Female | 55 | 180102cgf7 | 02/01/2018 1:... | T2 Diabetes... | 02/01/2018 | CBC Diff Profi... | MCV | 76.1 | (blank) | Loss apetitenausea... |
| 5 | HARMONAL | 00062196 | Female | 55 | 180102cgf7 | 02/01/2018 1:... | T2 Diabetes... | 02/01/2018 | CBC Diff Profi... | Monocytes | 9 | (blank) | Loss apetitenausea... |

**Figure 26.** De-normalized Meta Datasheet view of 3650 instances with defined variables as in Rule 2.




| Model | Accuracy | Runtime |
|---|---|---|
| Naive Bayes | 100.0% | 87 ms |
| Generalized Linear Model | 91.9% | 444 ms |
| Deep Learning | 100.0% | 23 s |
| Decision Tree | 73.8% | 398 ms |
| Random Forest | 100.0% | 14 s |
| Gradient Boosted Trees | 100.0% | 2 min 44 s |

**Figure 27.** Performance of Classification results for Rule 2 by applying known algorithms.

In a relatively larger dataset of 15696 instances of 100 endocrine patients with multiple visits and diagnoses, the RapidMiner Student platform was tested to give patient-specific results with much better speed than in Orange. Custom rules were applied to see relations in different parameters comprising 11 features out of 19 most relevant to make a diagnostic decision were extracted features in a given Dataset 3 (Table 1). Decision Tree was found best in accuracy 56.4% in a minimum time of 5s. The second best was deep learning analytics which kept balance in accuracy of 55.3% and speed to 24s. Results were recorded for DM Diagnosis for a 42% Confidence Interval with most likely Comorbidity Diseases. Errors were witnessed due to small class data in various analytical models implying that it would perform even better with big data. Deep Learning analytics showed in Figure 28 a male patient of age 54 diagnosed with DM with several comorbidity diseases. On validation, the tests and results inputs contradict the diagnosis, which explains that the actual diagnosis for the female patient, aged 49, was DM. The comorbidities found are; low back pain only. It is necessary to validate the results with actual diagnosis. In the case of a new patient, confirming a doctor is important as the analytics may predict the chances of occurrence of other comorbidity diseases as in Figure 28.

Results obtained from the decision tree model in Figure 29, were found interesting as they remained static for every patient with differing parameters. In comparison with other algorithms, RapidMiner software stated the decision tree model as best, but it is over-fitted as the results remain the same for any change in parameters we do. Explanation of results

found through different ML analytical models given in RapidMiner on validation were found to highly contradicting of the actual diagnoses which is not satisfactory. Still, the results showed deep learning mechanism promising in comparison to decision trees and other known algorithms in Figure 25 and 27, which were found over-fitted and highly inefficient in terms of result predictions as explained referring to Figure 29.

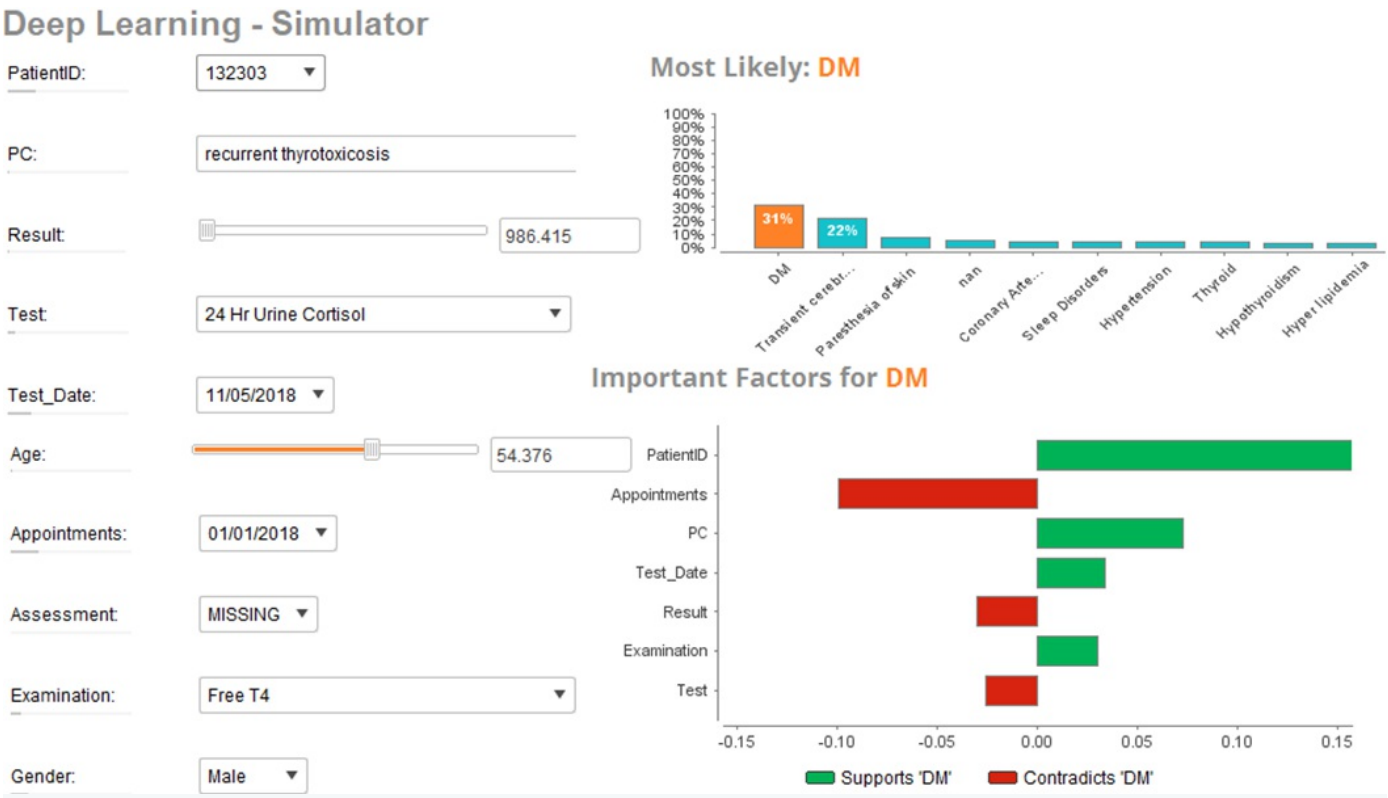



**Figure 28.** Patient is diagnosed DM with several comorbidity diseases using Deep Learning Analytics.

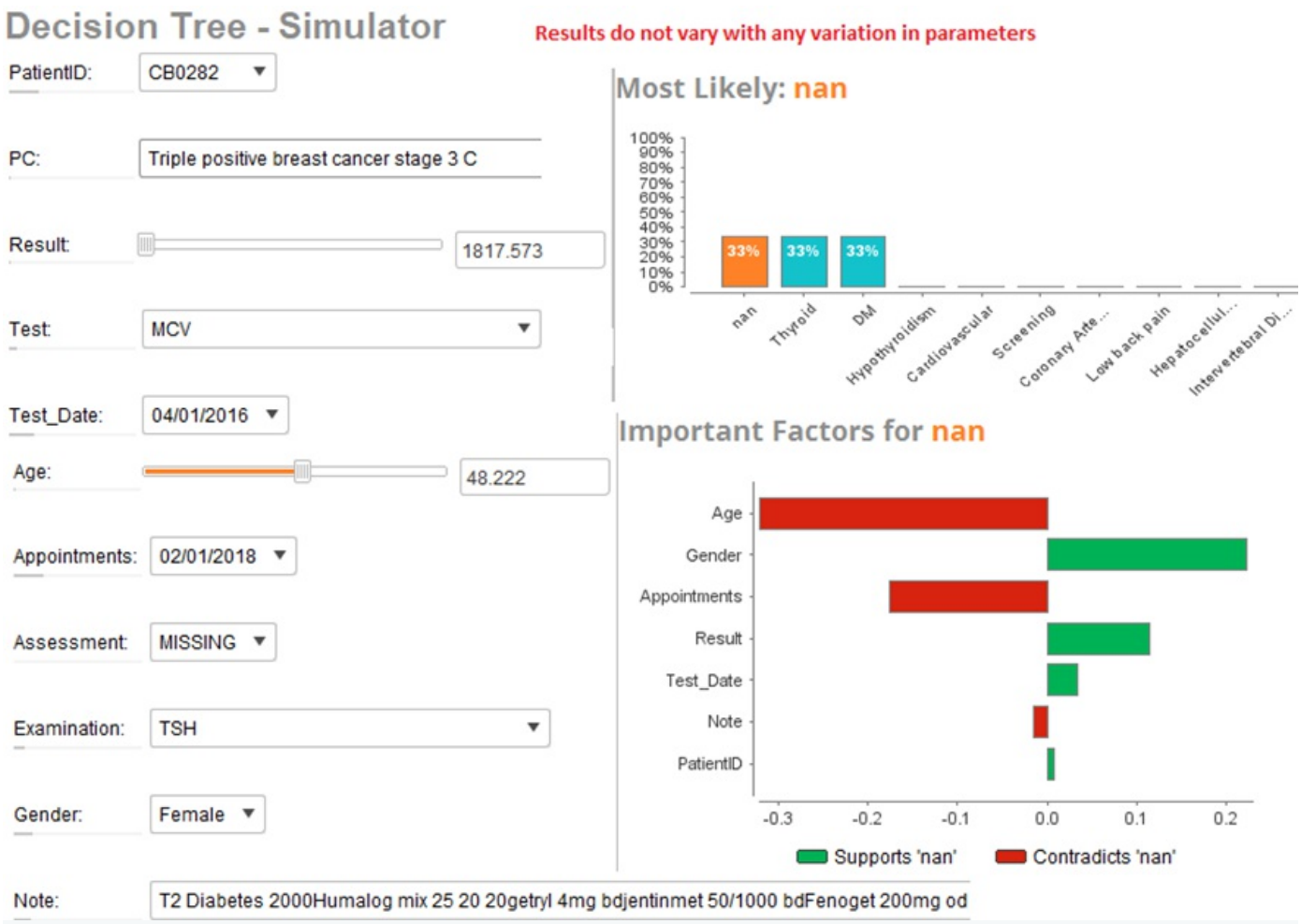



**Figure 29.** Decision Tree results when applied on 15696 records of endocrine patients (over fitted).

RapidMiner analytical models were then tested on comparatively big temporal data of 87803 instances of 14407 endocrine patients linked with a patient unique identifier, Visit No. (VAN), and date with target values of ‘Test’ and ‘Diagnoses’. There were 53592 missing labels in ‘Diagnoses’. Important 16 features from Dataset 3; VAN, VISIT_DATE, AGE, GENDER, EXAMINE, RESULT, ALLERGY, NOTE, ASSESSMENT, PC, ICD_CODE, REF_CODE1, MEDICINE, STRENGTH, UNIT_DESCRIPTION, DAYS, were taken with Diagnosis as target label and other special attributes were Patient-ID and Test to pass through analytics for giving different accuracy and speed results on validation of known ML algorithmic models as in Figure 30 to show deep learning performing its best on big data. Big data results were not patient specific as visualized in Figure 31, where deep learning was found to perform best and with temporal features the performance was even better.

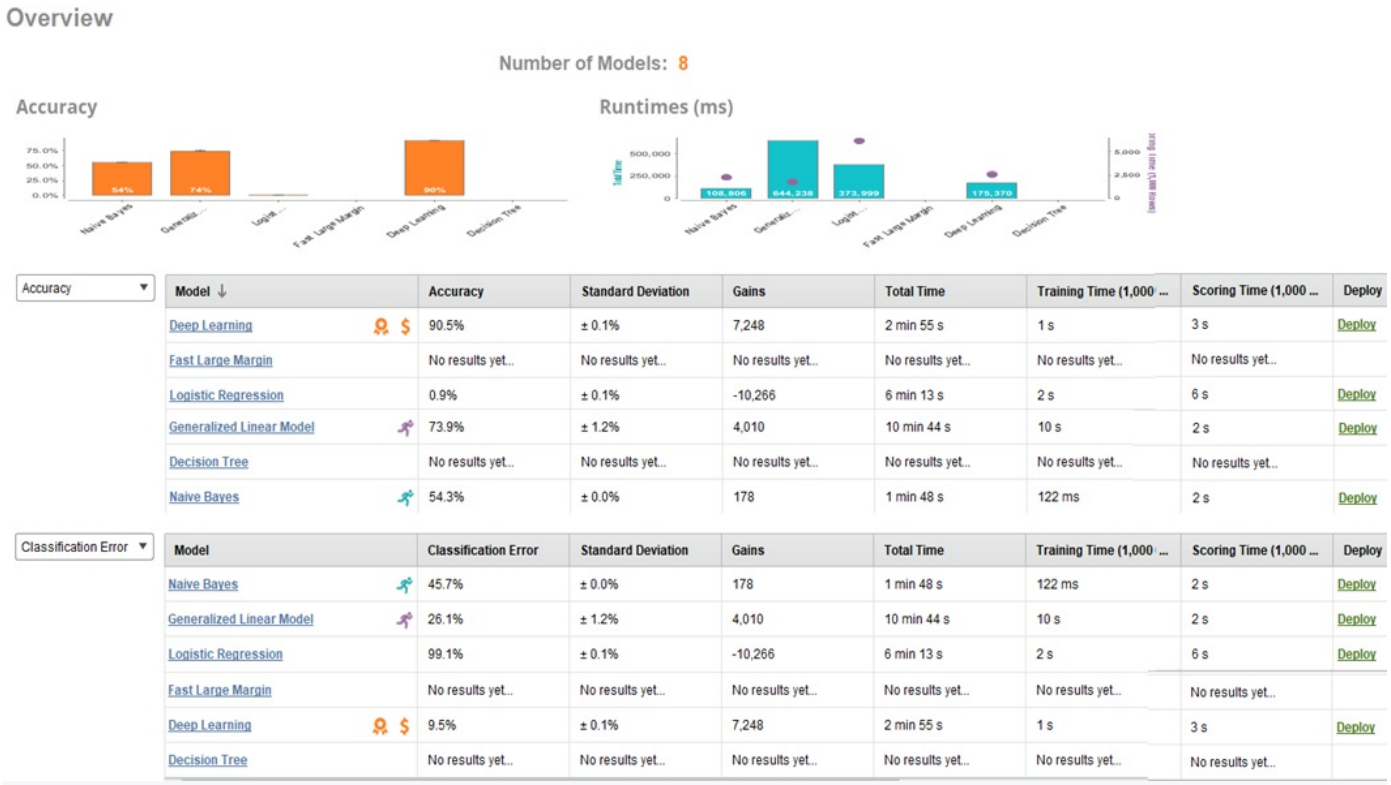



**Figure 30.** Deep Learning analytics found best to perform on temporal big data of 87803 records from 14407 endocrine patients with multiple visits.


**Figure 31.** DM diagnoses with other comorbidity diseases for patients male and female having NKDA (No Known Drug Allergy).

## 8. Insight on Experimental Results

Multiple experiments run in different settings and ML platforms with 3 endocrine datasets of variant sizes as in Table 1 and 2 with varying feature selection methods from phenotypes [50] gave varying performances in regard to speed and accuracy (Table 3).

CASE 1. Orange Framework with a hybrid unification of Louvain, Manifold and Hierarchical Clustering into Louvain Mani-

Hierarchical Fold Learning (LMHFL) was found very efficient on a small dataset of 3650 as well as 15696 instances from multiple visits of 100 endocrine patients. Results from 3650 records were patient-specific letting the user view the detailed diagnosis of each patient clustered as males and females. 15696 records took almost 48 hours to process. Maximum Correlations were found in features and rules were inducted with a maximum accuracy of 0.7 with Laplace where entropy measure failed for single target variable ‘Diagnosis’.

CASE 2. Qlik Sense is another cloud-based statistical tool to study and correlate datasets of different dimensions through normalizing. Insights gained from datasets of 100 and 14407 patients were found interesting to understand various relations for diabetes patients having various other comorbidity diseases as in Figure 18, 19, 20, 21 and 22 in section 7.

CASE 3. RapidMiner ML tool for education was used to evaluate various algorithms on given datasets in terms of speed and accuracy. Deep learning algorithmic automated model [14] was found to be fast and gave maximum accuracies of 96.7%, 100% and 90.5% on the three datasets of varying sizes (Table 1) (Figure 31, 27 and 30). It performed best on temporal data. The Decision Tree model on a given dataset was found over-fitted (Figure 29). Different patient-specific relational rules were extracted on the smaller dataset (Figure 28), given;

Patient_ID – Gender – Test – Test Date – Assessment -> Diagnosis of DM – Comorbidity Diseases, given a confidence level of 42% and its comorbidity diseases. In Figure 25, diagnosis is seen as related to allergy as well.

**Table 3.** Justification of variable sized datasets taken giving different results on different analytical platforms.

| | Dataset 1 | Dataset 2 | Dataset 3 |
|---|---|---|---|
| **RapidMiner** | Commercial platform gave us the capability to get higher accuracies of 100% with multiple labels with na¨ıve bayes (in 87ms), Random Forest<br><br>(in 14s), and deep<br><br>learning (in 23s) | Using educational platform, on larger dataset using 42% confidence level for<br><br>DM, same algorithms performance varied. Na¨ıve Bayes failed with 37.2% accuracy, Decision Tree perceived accuracy was 56.4% in 5s but when results were validated it was found to be over-fitted. Gradient boosted trees perceived accuracy was 56% and time took was more than 42mins. On validation results were not found informative. Random Forest was also found mostly inaccurate on validation and slow. Deep learning did better here with 56.4% accuracy and patient specific explainable results on validation. | It did equally well on much bigger dataset but was not patient specific. Deep learning was found to give max. accuracy of 90.5% in 2mins 55secs. Generalized Linear<br><br>Model gave accuracy of 73.9% but was slow. Na¨ıve Bayes processed faster than deep learning but perceived accuracy was much lower to 54.3% only. These results were get using temporal data like; Visit Date. On non-temporal data deep learning gave 62% accuracy in<br><br>2mins 9secs.<br><br>Then custom Deep Multinomial Model with multi-label classification was applied that was fast showing distributed deep learning capabilities. It processed all three datasets above 30k records.<br><br>The model performed best with log loss of 0.08<br><br>(Figure 36) 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 (Figure 37 and<br><br>38). |
| **Orange** | 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. | On a big dataset with 9646 records, 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. Took almost 2 days to process results. | It would not data over 10k records due to limited resource capacity.<br><br>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. Laplace gave max accuracy above 0.9. |
| **QlikSense** | 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 |

The deep learning algorithm also showed promising accuracy of 90.5% in only 2min 55sec on Big Data comprising 87803 rows and 16 columns (features) having ‘Diagnosis’ as the target variable with ‘Patient ID’ and ‘Test’ having higher weight

(Figure 30). Decision Tree failed to give results in time for big data with other known ML algorithms. The deep learning model was multinomial based on four layers with 10 epochs and 32-class classification. The total training sample size was 184,740 on mini-batch size equaling 1. Cross-Entropy or log loss was only 0.08126 or 0.0711 with a multi-label model that is near 0 showing close similarity between predicted and actual results. R-Squared was 0.99977 and MSE was 0.01686 only to show model is fit (Figure 10).

## 9. Evaluating Analytical Algorithms

Researchers are finding deep learning algorithms give high performance in speed for big or more complex data. Famous are convolutional neural networks (CNNs), deep residual learning (ResNet), capsule networks and multinomial logistic regression models [51]. These models have shown great performance on the MNIST dataset. CNN is known to recognize patterns in images. ResNet also uses deep convolution operation, but it differs in structure and depth and can go up to 100 connected layers and beyond. Capsule networks keep records of positions of inputs while in CNN all inputs are pooled and positions are often lost. Capsule net is able to generalize from a small amount of data for training. Multinomial logistic regression (MLR) acts as an alternative to naïve Bayes and is designed to do multiclass classification [52]. All these algorithms need high-performance computing resources to run big data like ours. Otherwise, a simple Jupyter Notebook using Python is not enough as shown here when we run Decision Tree and Multilayer Perceptron. The computing resources utilized were processor; Intel(R) Core(TM) m3-7th Generation CPU speed of 1.00GHz to 1.61 GHz, and RAM equaling 8GB. Therefore, Rapid Miner was used to test all known high-performance machine learning algorithms. Results were shown in Figures 25 to 31 for using Rapid Miner and deep learning was found better. Therefore, we run our hybrid analytics models; Deep Multinomial Distribution model and LMHFL in open-source cloud platforms like; Rapid Miner and Orange.

### 9.1. Decision Tree

The Decision Tree algorithm given by Lei Mao, in 2017, at the University of Chicago, was tested on our datasets of 100 endocrine patients. This algorithm classifies both categorical and numerical features with speed. Each categorical feature splits into branches for individual values without assuming. Numerical features are only split into two branches and numerical features may repeat themselves in the tree.

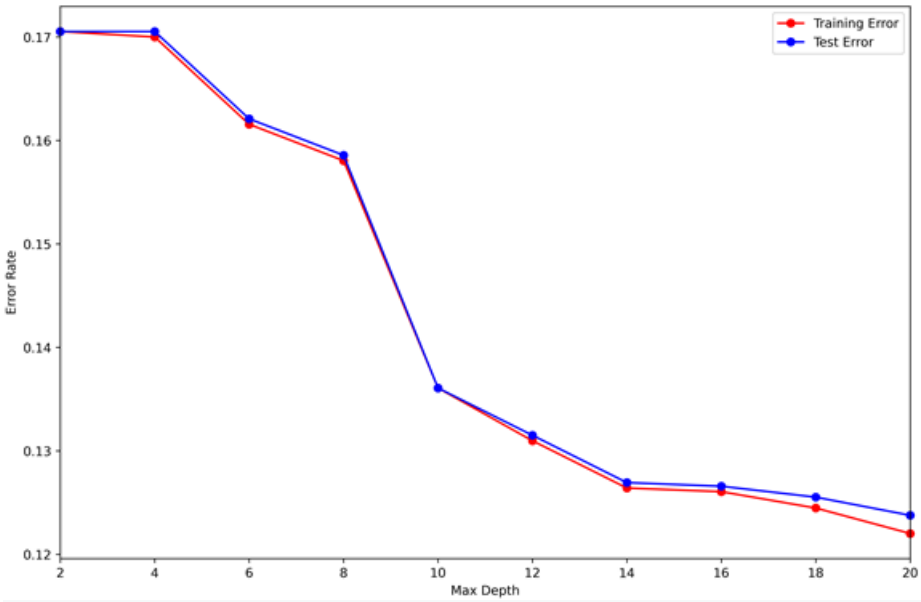



**Figure 32.** Validation Curve for Decision Tree algorithm run on 3650 instances with test error equaling to 0.124.

The algorithm ran successfully in Python in Jupyter Notebook, with 3650 instances. Tree max depth was 20 and the time elapsed was 1 minute 19 seconds only with a test error of 0.124 (Figure 32).

The algorithm was again run on 15696 instances but gave out of-bound index error due to lack of computing resources as no open-source cloud platform was used.

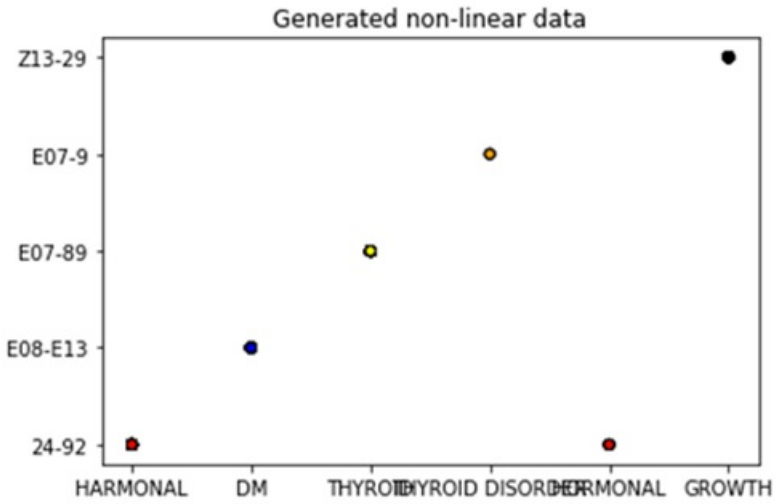
## 9.2. Multilayer Perceptron

Figure 3 in [53] shows different deep learning architectural models. Each referenced model has its connection structure given in terms of its node types as; input, hidden and output nodes. Multilayer Perceptron (MLP) is a type of ANN where every node in layer (l) is connected to each node in the next layer l+1, thus making it unidirectional with few hidden layers, that is a weighted sum of outputs of the previous layer (l) with an activation ($\alpha$) parameter for next layer. This nonlinear activation function normally takes sigmoid or tanh but now, it also takes rectified linear functions termed as ReLU.

Multilayer Perceptron runs in Jupyter Notebook using Python, giving visualization on 3650 instances with only two features ‘Diagnosis’, ‘ICD-10-CM’ and a target class ‘Diagnosed’ (Figure 33).

Rule: ‘Diagnosis’ – ‘ICD-10-CM’ → ‘Diagnosed’

Multilayer Perceptron failed to run on non-scaler and bigger datasets, giving the error of ‘too many indices’ for datasets with multiple labels and two or more features. The algorithm required a scaler dataset and, therefore, gave the value error ‘could not convert string to float’.




```
X_train: (2054, 2), y_train: (2054,)
X_val: (363, 2), y_val: (363,)
X_test: (427, 2), y_test: (427,)
Sample point: ['DM' 'E08-E13'] → DM
Classes: {'HORMONAL': 198, 'DM': 2359, 'THYROID': 248, 'THYROID DISORDER': 3, 'GROWTH': 36}
```

**Figure 33.** Data visualization to split into training, validation and test sets of 70%, 15% and 15% respectively.

## 9.3. Deep Learning Heuristics for Hybrid Analytical Model

Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN)[53] also run on scaler datasets for logical reasoning and need GPU to run using TensorFlow Keras. CNN mostly runs on images to extract meaningful information through convolutional filters aggregating extracted features. Where, RNNs are used for sequential data having time-series or natural language. The input nodes are activated not just from the current but also from previously hidden layers. Known deep learning algorithms failed due to a lack of computing resources and our experiments with cloud analytics on our datasets in sections 5 and 6 gave a clearer picture of implementing our hybrid approach for deep learning heuristics LMHFL as we used in Orange Framework in case 1.

LMHFL Model [48] as the name suggests is a hybridized deep learning heuristic using Louvain clustering, manifold learning and hierarchical classification for the best visualization of results [54]. Fair clusters with maximized edges in each cluster are possible using Louvain clustering [55]. All nodes are first put in different communities and then through greedy optimization strategy achieve maximum modularity to assemble all communities [49]. It lets us decide on the components and k number of clusters needed as hyperparameters. Manifold Learning [56] gave us the choice to use and compare multiple ML models like; t-SNE [57], local linear embedding [58], spectral embedding [59], Isomap [60], MDS, etc. in combination with Louvain and hierarchical clustering for best visualizations as multidimensional scaling (MDS) [61] and scatter plots filtering the outliers to get induced rules with maximum correlations.

We used Louvain Clustering with Manhattan followed by T-distributed stochastic embedding (t-SNE) that showed maximum accuracy in comparison to other methods with Chebyshev [62] metrics embedded in Manifold Learning to sharpen the features characterization for better visualizations. Patients were grouped for diagnosed DM and related

comorbidity diseases as in Figure 6 and 34. Associations of DM with its comorbidity diseases were visualized via MDS with a maximum correlation of 0.989 (Figure 6). Euclidean metrics were then used for weighted hierarchical visualization with maximum depth = 20. The scatter plot was again visualized, and outliers were filtered using a covariance estimator. The inliers data gave various distributions for visualizing the associations of features PatientID, TestID and Result with the Disease Diagnosed class (Figure 13). Rules were then induced from the CN2 Rule Induction algorithm with a maximum diagnostic accuracy of 0.506 up to 0.9 in different parameter settings through Laplace where the Entropy measure failed. Due to the lack of resources and speed of analytics to get the results for complete dataset 2 having all patients' profiles was slowed down to 48 hours for 15696 instances.

Both versions of LMHFL were then tested on individual patient profiles. In one case, an individual female patient diagnosed with DM and breast cancer as comorbidity resulted in induced rules, (Figure 35), using the Laplace metrics. The maximum quality achieved was 0.859, with an F1-score of 0.703 with the LMHFL model. The same patient profile when passed through Fast-LMHFL that integrated fast.ai DL library did not show quite good quality in induced rules and the F1-measure was only 0.527 still the confusion matrix showed better predictions compared to the LMHFL model. The lack of accuracy in Fast-LMHFL for induced rules and F1-score is associated with the missing text in the 'Note' and 'PC' fields.

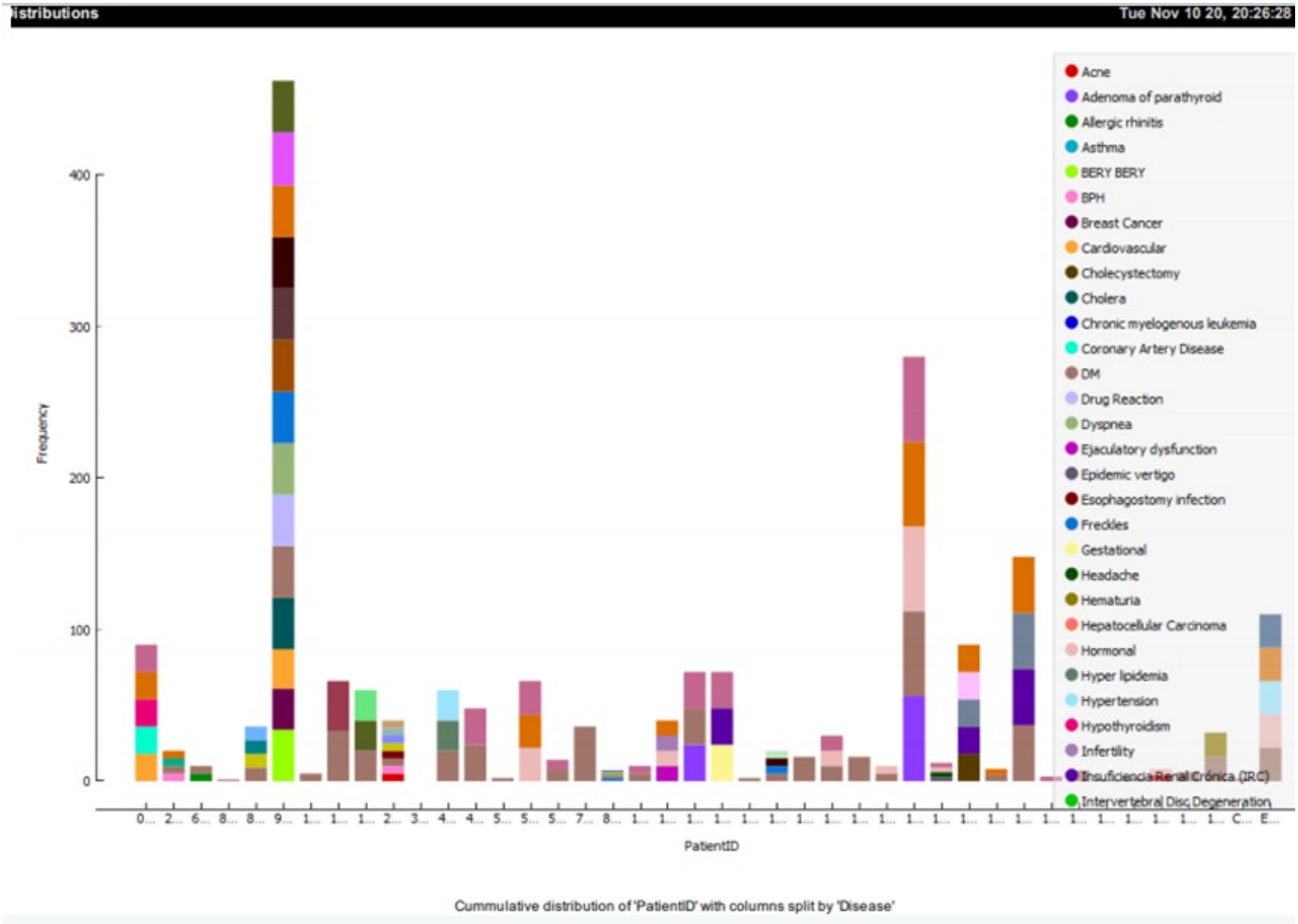



**Figure 34.** 100 Patients having DM and other associated diseases from dataset 2 of 15696 records.

**Data domain**

**Features:** C0, C1, C2, C3, C4, C5, C6
**Meta attributes:** ICD-10-CM, Cluster, mds-x, mds-y, Group
**Target:** Diagnosed

**Rule induction algorithm**

**Rule ordering:** ordered
**Covering algorithm:** weighted
**Gamma:** 0.7
**Evaluation measure:** laplace
**Beam width:** 5
**Minimum rule coverage:** 1
**Maximum rule length:** 11
**Default alpha:** 1.0
**Parent alpha:** 1.0

**Induced rules**

| | IF conditions | | THEN class | Distribution | Probabilities [%] | Quality | Length |
|---|---|---|---|---|---|---|---|
| | C6≥-8.193819945705201e-18 AND C3≤3.1220369779236196e-16 AND C1≥9.226236431288941e-17 AND C6≤-3.679172505787401e-18 AND C2≥-3.756042872991426e-17 AND C3≤-7.7342027775355004e-18 AND C0≥2.1801445163892007e-17 AND C6≥-7.089181884921548e-18 AND C3≥-4.436393782607856e-16 | → | Diagnosed=CA BREAST | [0, 78, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0] | 1 : 86 : 1 : 2 : 1 : 1 : 1 : 1 : 1 : 1 : 1 : 1 | 0.859 | 9 |
| 1 | C6≥-8.193819945705201e-18 AND C6≤-3.279599580520664e-18 AND C1≥9.226236431288941e-17 AND C1≤2.1909668331136466e-16 AND C1≤1.3874223245771494e-16 AND C1≤1.3124604944780613e-16 AND C5≥2.2356891558438564e-16 AND C6≤-4.10507472729613e-18 AND C1≤1.2670912244993433e-16 AND C5≤3.1704114065365297e-16 AND C3≥-3.171806033338691e-16 | → | Diagnosed=CA BREAST | [0.0, 107.5, 1.0, 0.0, 0.0, 2.0, 1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0] | 1 : 86 : 2 : 1 : 1 : 2 : 2 : 1 : 1 : 1 : 2 : 2 | 0.858 | 11 |
| 2 | C6≥-8.193819945705201e-18 AND C6≤-3.279599580520664e-18 AND C1≥9.226236431288941e-17 AND C3≤3.1220369779236196e-16 AND C3≤-7.734202777355004e-18 AND C6≤-3.6621031262697 51e-18 AND C0≤2.1126401356567794e-17 AND C3≥-3.096391275725147e-16 AND C1≥9.877601232674215e-17 AND C6≥-7.748072219666585e-18 | → | Diagnosed=CA BREAST | [0.0, 148.3, 0.0, 2.0, 0.0, 2.7, 1.0, 0.0, 3.0, 0.0, 0.0, 2.0, 1.7] | 1 : 86 : 1 : 2 : 1 : 2 : 1 : 1 : 2 : 1 : 1 : 2 | 0.860 | 10 |
| 3 | C1≤-0.3221559087693022 AND C1≥-0.8207212796600066 AND C6≥-1.0368784155281014 AND C6≤1.6722597969051018 AND C1≥-0.8207212796600063 AND C5≥-0.4277643029969104 AND C5≥-0.4277643029969104 AND C3≥-0.968173132283626 AND C5≥-0.4277643029969103 AND C5≥-0.3138901515785705 3 | → | Diagnosed=CA BREAST | [7.0, 472.0, 16.0, 7.0, 0.0, 35.0, 9.0, 0.0, 8.0, 9.0, 4.0, 9.0, 6.0] | 1 : 79 : 3 : 1 : 0 : 6 : 2 : 0 : 2 : 2 : 1 : 2 : 1 | 0.795 | 10 |
| 4 | C0≥1.7244899464492837e-17 AND C2≥-6.804138443627991e-17 AND C1≥9.226236431288941e-17 AND C5≤4.4266715153312 45e-16 AND C3≤-2.1445243449412905e-17 AND C6≤-3.679172505787401e-18 AND C2≥-6.499389359656644e-17 AND C6≥-8.700503820422633e-18 AND C2≥-4.662789603203095e-17 AND C3≤-2.278153943508124e-16 AND C5≤2.5298643617840224e-16 | → | Diagnosed=CA BREAST | [0.0, 130.1, 3.0, 1.0, 0.0, 1.7, 4.0, 0.0, 0.7, 1.0, 0.0, 1.7, 0.0] | 1 : 84 : 3 : 1 : 1 : 2 : 3 : 1 : 1 : 1 : 2 : 1 | 0.839 | 11 |
| 5 | C1≤-0.3043851492997129 AND C1≤-0.3221559087693022 AND C1≥-0.8207212796600066 AND C3≥-0.968173132283626 AND C6≥-1.0368784155281014 | → | Diagnosed=CA BREAST | [6.9, 447.4, 15.2, 6.9, 0.0, 32.5, 7.3, 0.0, 7.6, 8.3, 4.8, 8.3, 8.2] | 1 : 79 : 3 : 1 : 0 : 6 : 1 : 0 : 2 : 2 : 1 : 2 : 2 | 0.792 | 5 |
| 6 | C6≥-9.312936604785421e-18 AND C3≤1.0909794396158525e-16 AND C1≥9.226236431288941e-17 AND C6≤-3.679172505787401e-18 AND C6≥-8.631392669795324e-18 AND C0≤2.1126401356567794e-17 AND C4≤2.247437752408937e-16 AND C3≥-3.438723026394813e-16 AND C1≤1.6197877736559407e-16 AND | → | Diagnosed=CA BREAST | [0.0, 97.6, 0.0, 0.7, 0.0, 1.2, 0.0, 0.0, 0.7, | 1 : 85 : 1 : 1 : 1 : 2 : 1 : 1 : 1 : 1 : 3 : | 0.846 | 11 |

**Figure 35.** Induced Rules from extracted female patient profile of age 44 from Dataset 2 for DM and its comorbidities with accuracy=0.859

This LMHFL model proposed in [48] was then tried to analyze each individual patient profile from a big dataset of 14407 endocrine patients. This dataset was found limited as each patient was diagnosed with one disease either it be DM or its comorbidity. The rules were induced for Practitioners Comments (PC) and Notes with Tests, Allergies or Medicine as Target label classes related to these diseases which acted here as a constant variable. Example cases analyzed were patients diagnosed with DM in 2007 and then went for a Liver Transplant in 2018, a patient diagnosed with Thyroid is on insulin and notes indicate that he also suffers from hyperglycemia and another case diagnosed as Steroid Induced Diabetes having symptoms like dry mouth and ankle swelling who suffered from known metastatic carcinoma of ovary started on Solu-Medrol and then Deltacortil steroid (used to treat severe allergic reactions) and her BG was 400-500. All these patients were identified as having no known food or drug allergies (NKFA, NKDA or both). Researchers confirmed the disease to be a comorbidity of DM based on tests conducted like; HbA1c and medicine taken with Lantus insulin for example. Laplace as usual won over Entropy and WRAcc for evaluation measures keeping the accuracy between 0.7 and beyond 0.9 in some scenarios. However, these differences in accuracy results create bias within entropy and Laplace measures. While Laplace shows maximum accuracy, entropy measures show an accuracy close to zero. Communities were made using Louvain and further spectral embedding and Isomap were applied within Manifold Learning for better results. Multidimensional and Scatter plots were again visualized through results from hierarchical clusters.

### 9.4. Deep Learning Multinomial Model

RapidMiner commercial model tested on the trial version was found to be slow on a large dataset, therefore, an educational platform specifically designed for researchers was chosen and discussed in case 3 of section 5. It gave researchers the capability to study advanced ML models for getting highly effective and accurate results classifying huge datasets. Famous are logistic regression, Naïve Bayes, Generalized Linear Model, Decision Trees and multinomial logistic regression models [51]. In our experiment, the Multinomial Logistic Regression (MLR) Deep Learning Model gave an accuracy of 90.5 on our 14407 patient Dataset 3 of 87803 records. MLR showed higher accuracy than naïve bayes in our experiment (Figure 30) due to its ability of multiclass classification and for this reason, the speed becomes slightly lower.

We then developed a custom model with the same deep-learning multinomial model and cross-validated it with a decision tree. The model took a rule set having 17 features that were given different target roles; Patient_ID as batch, Diagnosed was categorized into three roles as target label class, predicted and cluster, ICD_Code was also set as target label class, other features that were set for prediction were Test and Allergy, other than Diagnosed there were other three features set to form clusters namely Age, Gender and Examine, with other features set as regular and that were Visit_Date, Note, PC, Result, Medicine, Unit_Description, Days and Strength and finally VAN was set as id as an identifier for unique visit account no. for each patient. Deep learning multinomial model used tanhdropout and softmax as activation functions. L1 and L2 regularizers within the model assure that well-fitted results are generated in H2O 3.30.0.1 that uses multi-layer feed-forward ANN trained on stochastic gradient descent with backpropagation. The model performed best with a log loss of 0.08 only (Figure 36) with two hidden layers of size 50 which were tunable for more accurate results.

This balanced trained model then cross-validated its results 10-fold using a decision tree that gave 100% recall and precision results with kappa equaling 1 (Figure 37). Two target labels used as a rule set; ‘Diagnosed’ and ‘ICD-10’ were the reason for multi-label classification using multinomial distribution present in the deep learning model.

Further, for a more accurate rule set and better classification of DM and its comorbidities the model was hybridized with a multi-label ensemble classification model where all 16 features were set as target labels leaving behind ‘Patient_ID’ that was set as batch parameter in deep multinomial model that was still the input here as resultant prediction model.

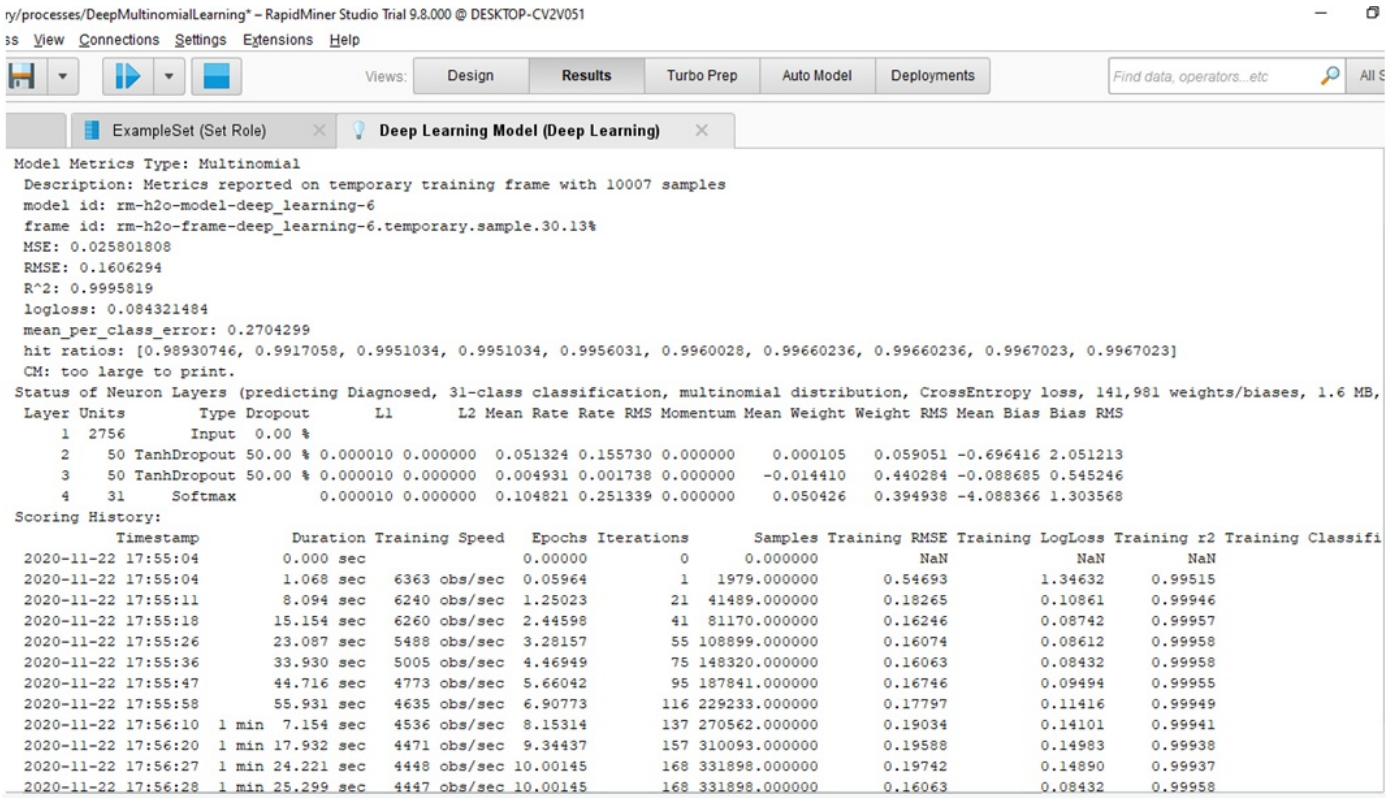
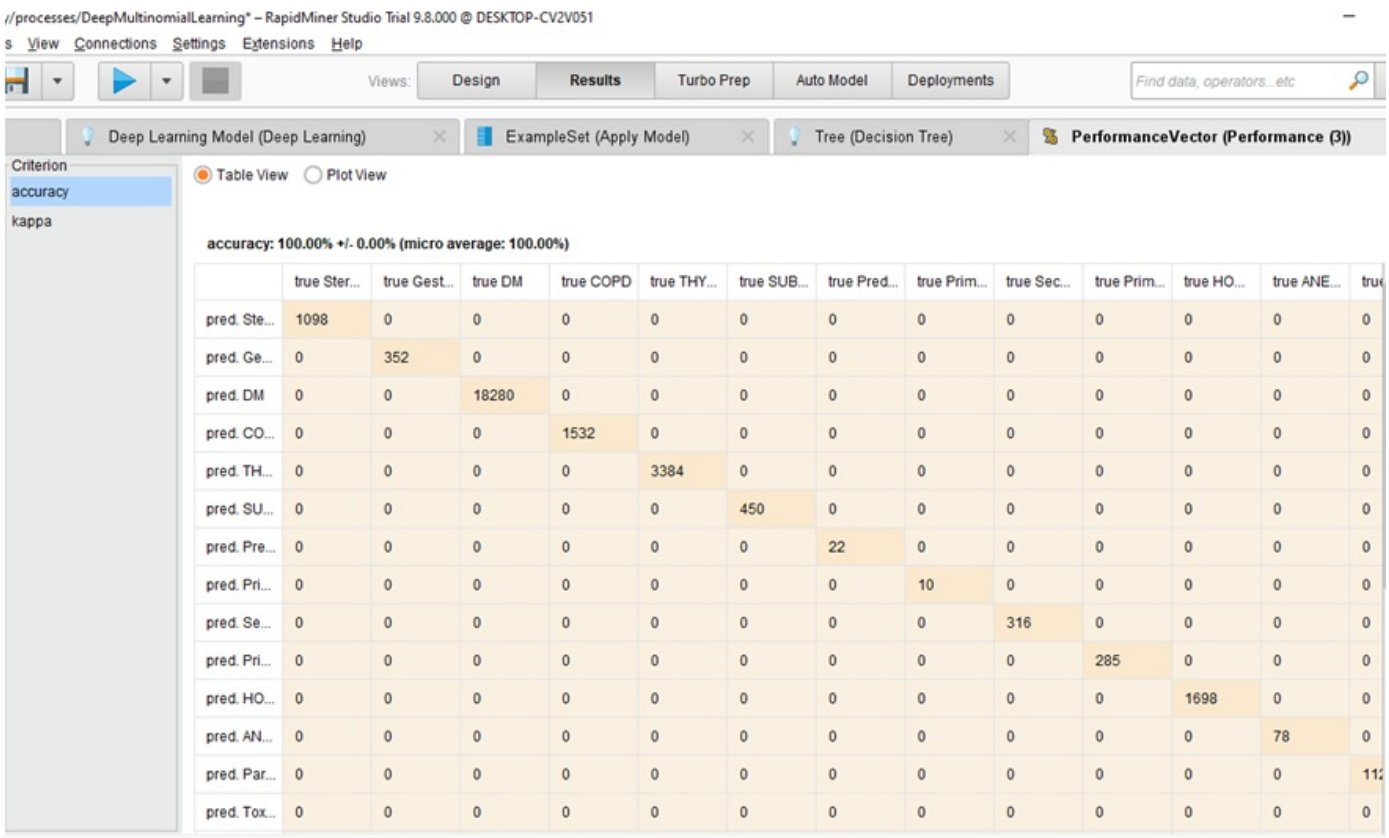



**Figure 36.** Descriptive accuracy results with log loss equaling 0.08 on 2 hidden layers of size 50 each.

This hybrid model showed slightly better accuracy results having a log loss of 0.202 (Figure 52 in Section 13) or further tuned to 0.0765 with hidden layers tuned to sizes of [31, 36] or [50, 50] respectively, for a multiclass trained model using different target labels to explore precise and complete rule sets for diagnosis of diabetes and its comorbidities. Log Loss nearer to 0 signifies a better and robust model. Finally, it was again validated with a split ratio of 0.7 and 0.3 for training and test sets respectively using a decision tree resulting in 100% precision and recall as in Figures 37 and 38.




| | true Ster... | true Gest... | true DM | true COPD | true THY... | true SUB... | true Pred... | true Prim... | true Sec... | true Prim... | true HO... | true ANE... | tru |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 | 11 |
| pred. Tox... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Figure 37.** Confusion Matrix with 100% precision and recall on 10-fold cross-validation for Deep Multinomial Learning Model.

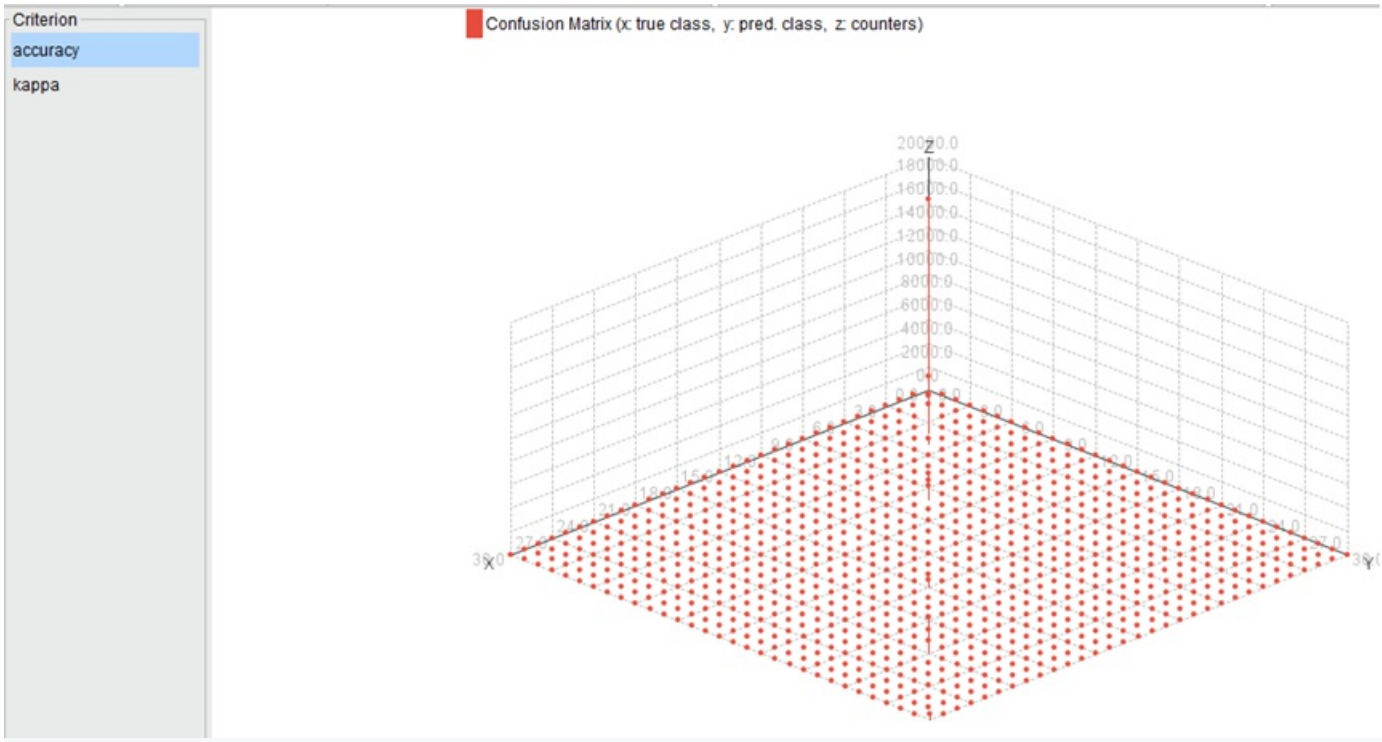



**Figure 38.** Box plot of Confusion Matrix with 100% precision and recall through split-validation for Multi-Label Deep Multinomial Learning Model.

## 10. Individual DM Patients Analytics for Diagnosis and Prognosis of Comorbidities

The 100 patients' dataset having 15696 instances that were left with 9304 records after pruning was filtered for 20 individual DM patient chunks having multiple diagnoses with comorbidity diseases mentioned in Table 2. 3 patients were first tested and validated for DM diagnosis with other associated comorbidity diseases; one male patient profile of age 53 yields 24 instances having DM and other diseases like Benign Prostatic Hyperplasia (BPH), and Premature Ejaculation (PME), the second patient is female with age 49 having 144 instances with DM diagnosis and Lower Back Pain (LBP) as a comorbidity disease, and the third patient is again a 73 years old male having 352 instances with DM and other comorbidities as Transient Ischemic Attack (TIA), BPH, Lower Urinary Tract Symptoms (LUTS), acne, sleep starter (Jerk), Paresthesias and Throat Infection. Each patient profile went through the LMHFL analytics model to test the accuracy of DM diagnosis and prognosis of associated comorbidity diseases in the form of MDS and scatter representations as clusters. Different model parameters were tried on different selections of features. Finally, Louvain with resolution ($\lambda$=1) and Isomap in Manifold and hierarchical clustering was chosen with distance metrics; Manhattan and Euclidean to run on uniform features set. 7 categorical features; gender, age, note, examination, test, result and PC, 1 Meta Attribute taken was ICD-10-CM to label with the target class, Diagnosed. The analytics model had other parameters like k-neighbors, and correlation parameters like Pearson or Spearman, as flexible to decide the best cluster formation with increased modularity and dimensionality reduction for visual representation in MDS, hierarchical tree distributions and scatter plots. The accuracy metrics were designed by visualizing induced rules using Laplace to weigh the associations within DM and its comorbidity diseases. Further, the model was validated using predictions and test accuracy resulting in area under the curve (AUC), classification accuracy (CA), F1 score, and precision and recall metrics. Finally, the confusion matrix shows the results of prediction accuracy and errors in the diagnosis of DM and its comorbidities if any.

Patients Set 1 (Pi) = {P1: 24, P2: 144, P3: 352}

Major representations are recorded as MDS plots in Figure 42, where we see associations between DM and its associated comorbidity diseases in three scenarios (patients). We also observed that the clusters formed are disjointed as the limitation found in Louvain as mentioned in [45]. This limitation posed by Louvain and visualized in the MDS plot was mitigated by connecting the graph in a tree structure using hierarchical clustering and thus it was visualized as a very balanced scalar output as the slope in a scatter plot for a non-linear and complex dataset (Figure 43). This scalar plot was possible with the selected feature set having only one Meta attribute of ICD-10-CM. There were several visualizations observed for different feature sets that had multiple Meta attributes for labeling the outcomes but the scatter plots were dispersed as in Figure 39, 40 and 41.

The selected features set for three patients in set 1 (Pi), had 7 variables, a single Meta attribute and a target class, which gave us a balanced connected scalar scatter plot, and was further noted for maximum Pearson correlations of 0.667, 0.475, and 0.986 for 24, 144 and 352 records of patients (Pi) respectively. The interpretation that we observed here is that patient suffering from a larger number of diseases has a higher correlation. Further evaluations were studied through Laplace quality measure for CN2 induced rules for each diagnosed disease depending on the weight of associations between the components or features ranging rule length from 2 to 11 maximum equaling to 0.8, 0.941 and 0.814 for each patient in the same order as referred to before as set 1 (Pi). The next evaluations done are through a comparison of

prediction and test accuracies in terms of AUC, CA, F1, precision and recall.

Prediction accuracies for Set 1 (Pi) = {AUC[0.986, 1, 0.974], CA[0.917, 0.993, 0.741], F1[0.911, 0.933, 0.714], Precision[0.926, 0.993, 0.820], Recall[0.917, 0.993, 0.741]}

Test accuracies for Set 1 (Pi) = {AUC[0.985, 1, 0.974], CA[0.913, 0.929, 0.731], F1[0.907, 0.925, 0.701], Precision[0.923, 0.935, 0.816], Recall[0.913, 0.929, 0.731]}

Finally, the confusion matrix in Figure 44, depicts clarity in results in terms of accurate predictions and a number of false predictions.

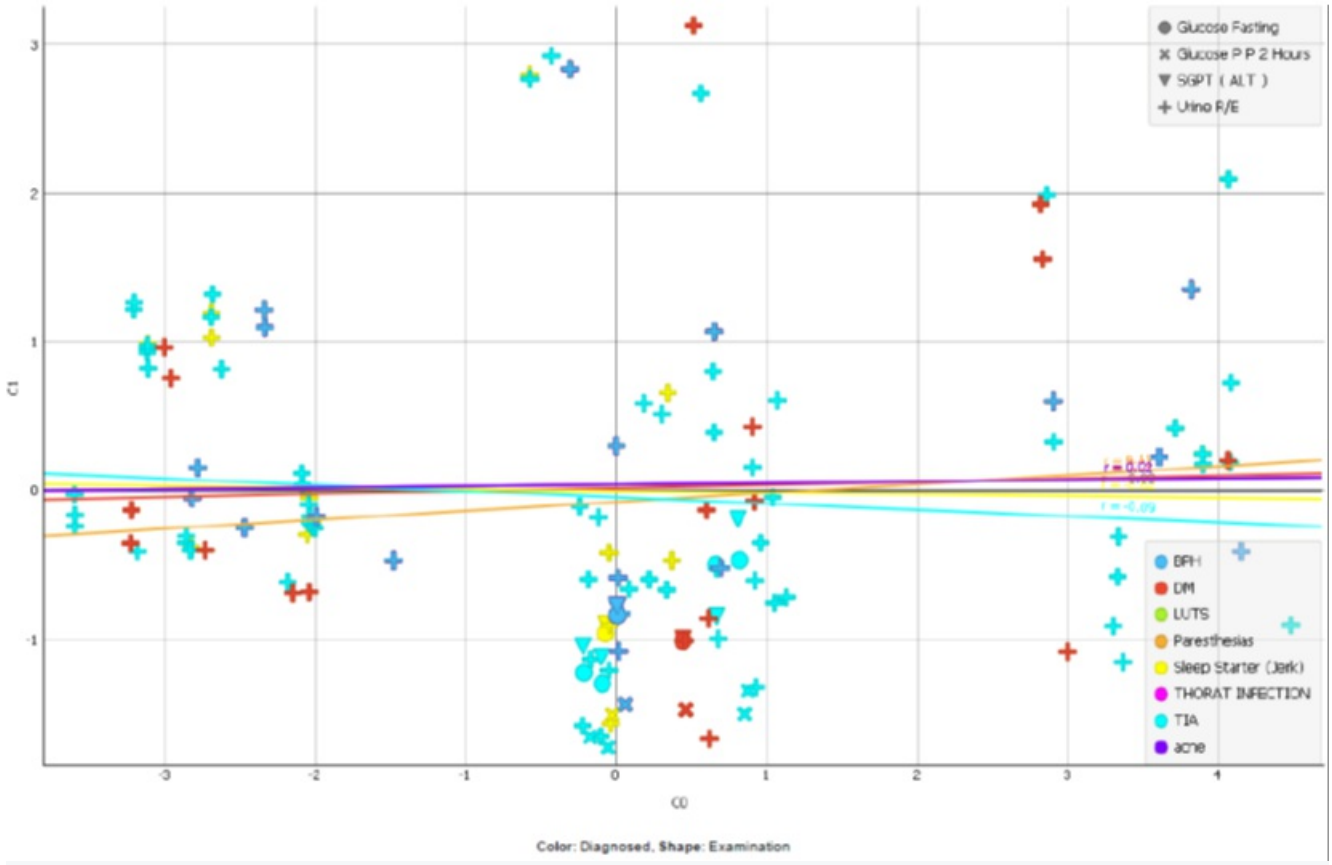



**Figure 39.** Multiple Meta Attributes generated Dispersed Scatter Plot showing association in DM and Comorbidities; Transient Ischemic Attack (TIA), BPH, Lower Urinary Tract Symptoms (LUTS), acne, sleep starter (Jerk), Paresthesias and Throat Infection.

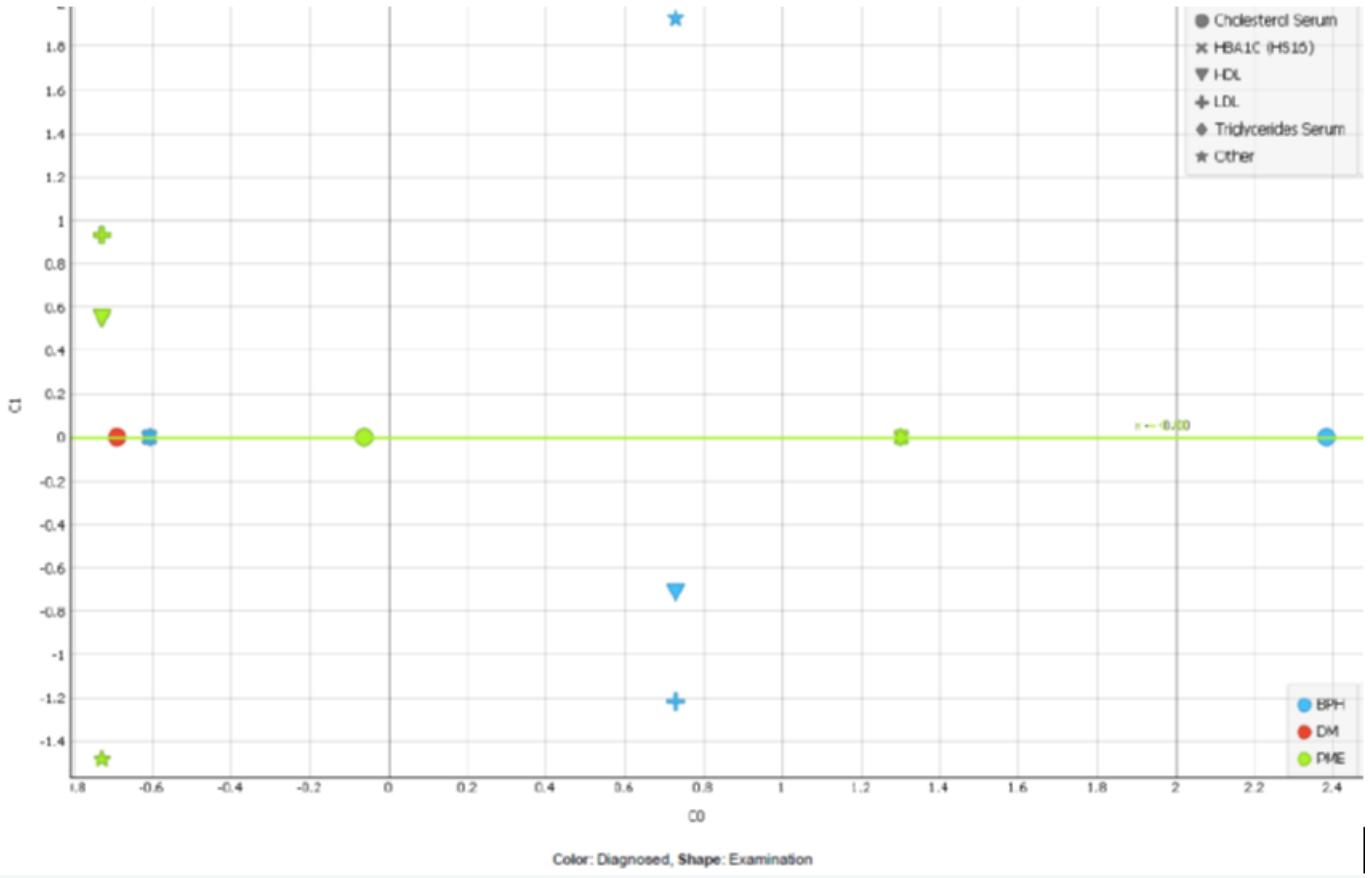



**Figure 40.** Multiple Meta Attributes generated Dispersed Scatter Plot showing association in DM and other two comorbidities; Benign Prostatic Hyperplasia (BPH), and Premature Ejaculation (PME).

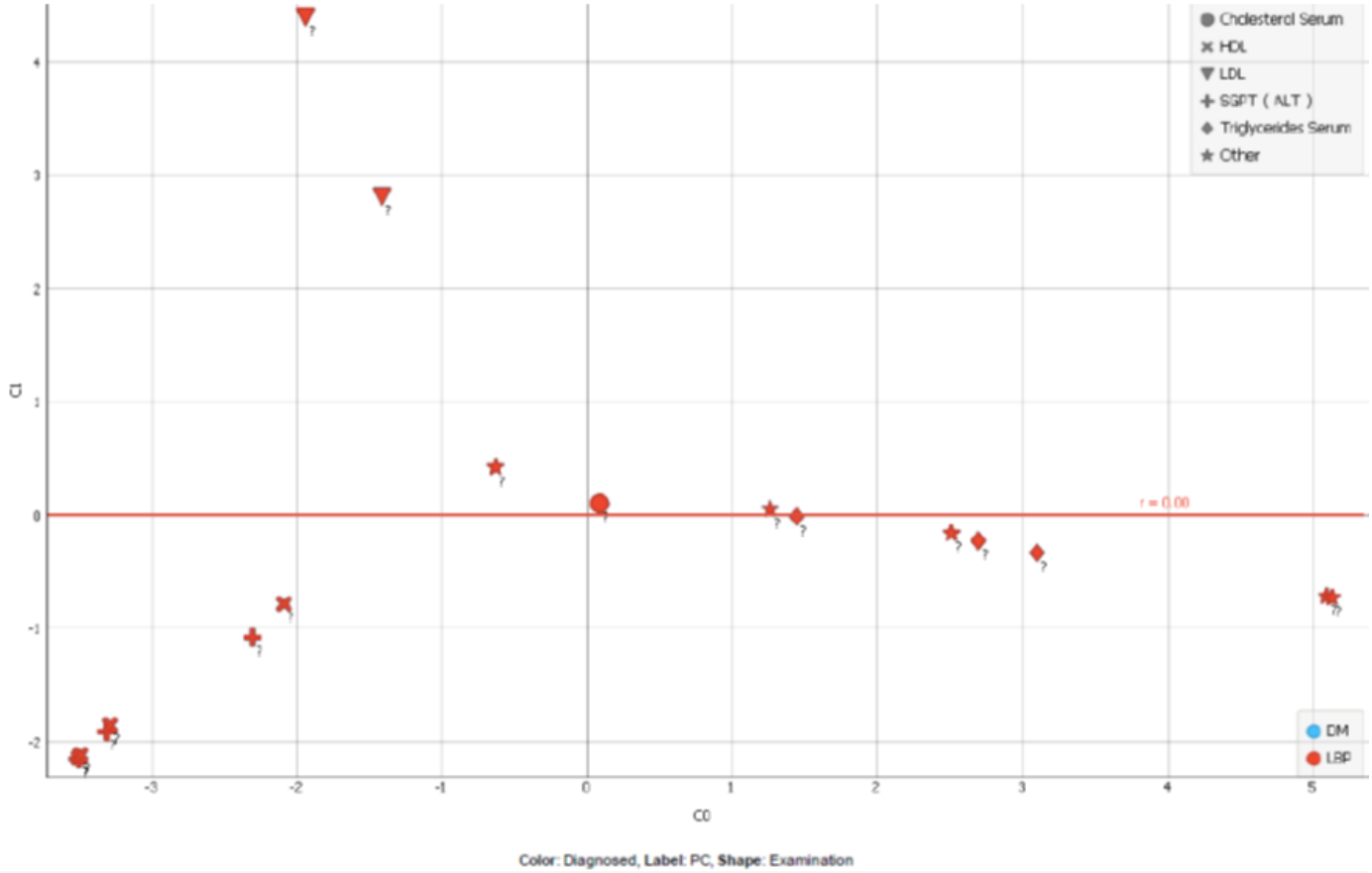



**Figure 41.** Multiple Meta Attributes generated Dispersed Scatter Plot showing association in DM and Lower Back Pain (LBP).

**Figure 42.** Multi-Dimensional Clusters plotted to show associations in DM and its comorbidities suffered by single patients with different phenotypes: a) Patient 1: Male, Age: 53, 24 records, DM diagnosed having two comorbidity diseases; Benign Prostatic Hyperplasia (BPH), and Premature Ejaculation (PME) b) Patient 2: Female, Age: 49, 144 records, Diagnosed: DM with Comorbidity: Lower Back Pain (LBP) c) Patient 3: Male, Age: 73, 352 records, Diagnosed DM and Comorbidities: Transient Ischemic Attack (TIA), BPH, Lower Urinary Tract Symptoms (LUTS), acne, sleep starter (Jerk), Paresthesias and Throat Infection.

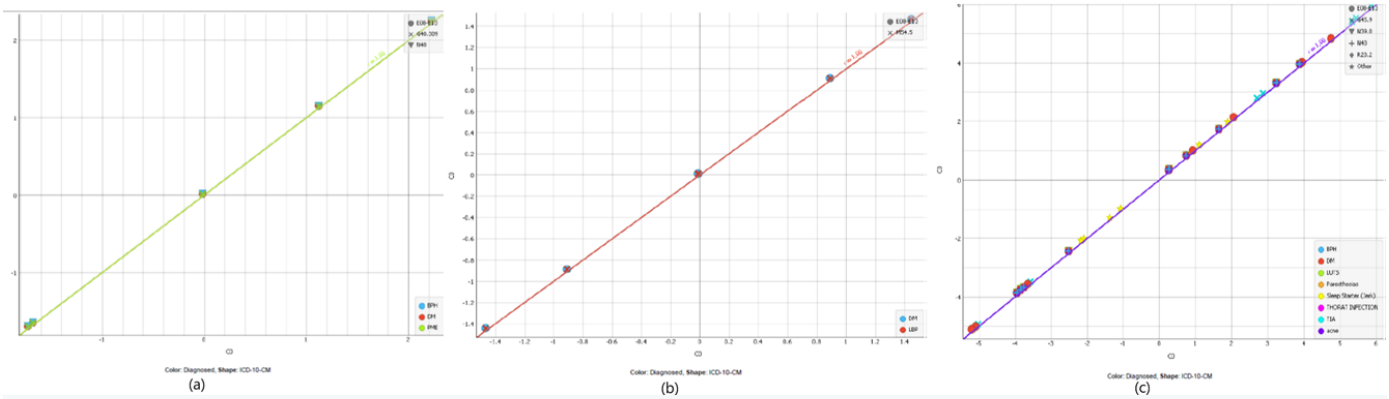
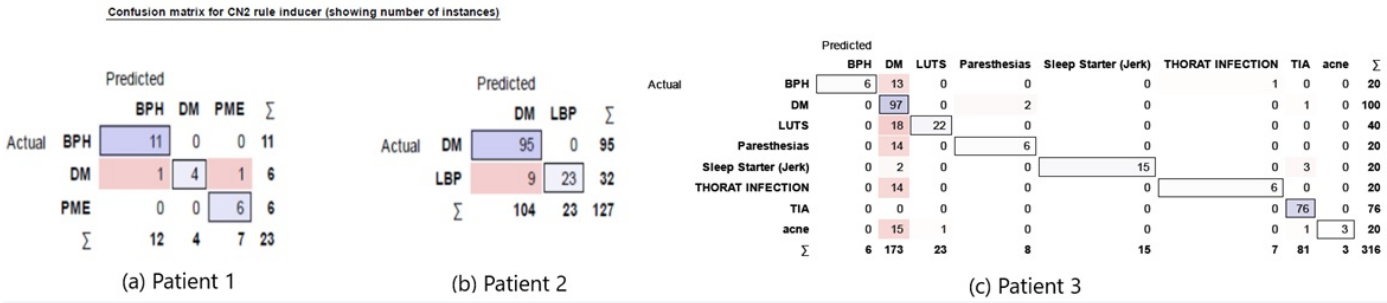


**Figure 43.** Scatter Plot to show a balanced slope for non-linear associations in DM and its comorbidities suffered by single patients with different phenotypes: a) Patient 1: Male, Age: 53, 24 records, DM diagnosed having two comorbidity diseases; Benign Prostatic Hyperplasia (BPH), and Premature Ejaculation (PME) b) Patient 2: Female, Age: 49, 144 records, Diagnosed: DM with Comorbidity: Lower Back Pain (LBP) c) Patient 3: Male, Age: 73, 352 records, Diagnosed DM and Comorbidities: Transient Ischemic Attack (TIA), BPH, Lower Urinary Tract Symptoms (LUTS), acne, sleep starter (Jerk), Paresthesias and Throat Infection.



Confusion matrix for CN2 rule inducer (showing number of instances)

| Actual \ Predicted | BPH | DM | PME | Σ |
|---|---|---|---|---|
| BPH | 11 | 0 | 0 | 11 |
| DM | 1 | 4 | 1 | 6 |
| PME | 0 | 0 | 6 | 6 |
| Σ | 12 | 4 | 7 | 23 |

(a) Patient 1

| Actual \ Predicted | DM | LBP | Σ |
|---|---|---|---|
| DM | 95 | 0 | 95 |
| LBP | 9 | 23 | 32 |
| Σ | 104 | 23 | 127 |

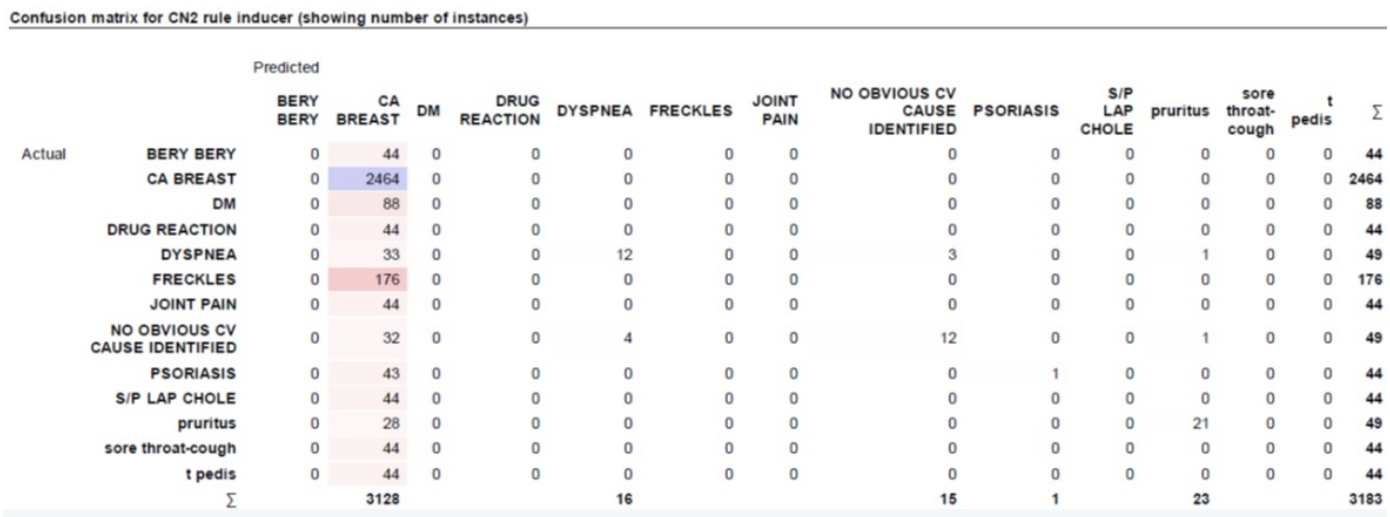
(b) Patient 2

| Actual \ Predicted | BPH | DM | LUTS | Paresthesias | Sleep Starter (Jerk) | THORAT INFECTION | TIA | acne | Σ |
|---|---|---|---|---|---|---|---|---|---|
| BPH | 6 | 13 | 0 | 0 | 0 | 1 | 0 | 0 | 20 |
| DM | 0 | 97 | 0 | 2 | 0 | 0 | 1 | 0 | 100 |
| LUTS | 0 | 18 | 22 | 0 | 0 | 0 | 0 | 0 | 40 |
| Paresthesias | 0 | 14 | 0 | 6 | 0 | 0 | 0 | 0 | 20 |
| Sleep Starter (Jerk) | 0 | 2 | 0 | 0 | 15 | 0 | 3 | 0 | 20 |
| THORAT INFECTION | 0 | 14 | 0 | 0 | 0 | 6 | 0 | 0 | 20 |
| TIA | 0 | 0 | 0 | 0 | 0 | 0 | 76 | 0 | 76 |
| acne | 0 | 15 | 1 | 0 | 0 | 0 | 1 | 3 | 20 |
| Σ | 6 | 173 | 23 | 8 | 15 | 7 | 81 | 3 | 316 |

(c) Patient 3

**Figure 44.** Confusion Matrix for three patients in set 1 (Pi) with actual and predicted results.

Another patient profile with 3528 records having breast cancer, Bery Bery, Dyspnea, Joint pain, freckles, Psoriasis, sore throat, t pedis and some other complications with DM was analyzed for tabular fields. It is understood from the inferences observed that due to the chronic condition of breast cancer, the model mostly overlapped all other diagnoses as seen in the confusion matrix in Figure 45.



Confusion matrix for CN2 rule inducer (showing number of instances)

| Actual \ Predicted | BERY BERY | CA BREAST | DM | DRUG REACTION | DYSPNEA | FRECKLES | JOINT PAIN | NO OBVIOUS CV CAUSE IDENTIFIED | PSORIASIS | S/P LAP CHOLE | pruritus | sore throat-cough | t pedis | Σ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BERY BERY | 0 | 44 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 44 |
| CA BREAST | 0 | 2464 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2464 |
| DM | 0 | 88 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 88 |
| DRUG REACTION | 0 | 44 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 44 |
| DYSPNEA | 0 | 33 | 0 | 0 | 12 | 0 | 0 | 3 | 0 | 0 | 1 | 0 | 0 | 49 |
| FRECKLES | 0 | 176 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 176 |
| JOINT PAIN | 0 | 44 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 44 |
| NO OBVIOUS CV CAUSE IDENTIFIED | 0 | 32 | 0 | 0 | 4 | 0 | 0 | 12 | 0 | 0 | 1 | 0 | 0 | 49 |
| PSORIASIS | 0 | 43 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 44 |
| S/P LAP CHOLE | 0 | 44 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 44 |
| pruritus | 0 | 28 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 21 | 0 | 0 | 49 |
| sore throat-cough | 0 | 44 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 44 |
| t pedis | 0 | 44 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 44 |
| Σ | | 3128 | | | 16 | | | 15 | 1 | | 23 | | | 3183 |

**Figure 45.** Female age 44 suffering from a chronic condition of DM with other underlying diseases, mainly breast cancer

**Figure 45.** Female age 44 suffering from a chronic condition of DM with other underlying diseases, mainly breast cancer.

## 11. FAST-LMHFL Diagnostics Model for Tabular Dataset Rich in Text Features

Individual patient profiles were analyzed in two model settings. In section 9, we have elaborated and discussed our results for analytics run on tabular dataset Set 1 ( ), where each feature was set as a category, Meta attribute was ICD-10-CM to label target class Diagnosis.

Here, the second scenario of the analytics model, Fast-LMHFL, is discussed, that was run on the same Set 1 ( ) with a tabular dataset and taking into account the text fields; PC and Note, for text mining. Patient 3 from set 1 ( ) only had text in PC and Note fields when converted to corpus in Figure 46. The query 'sleep' returned 88 records for concordance from Word Cloud fetched in Figure 47. The analytics were then run with Louvain clustering and the Manifold set for Isomap had 8 k-neighbors with components = 5 (the tabular features; gender, age, examination, test and result) and resolution=1 for best modularity clusters = 7 and reduced dimensionality.

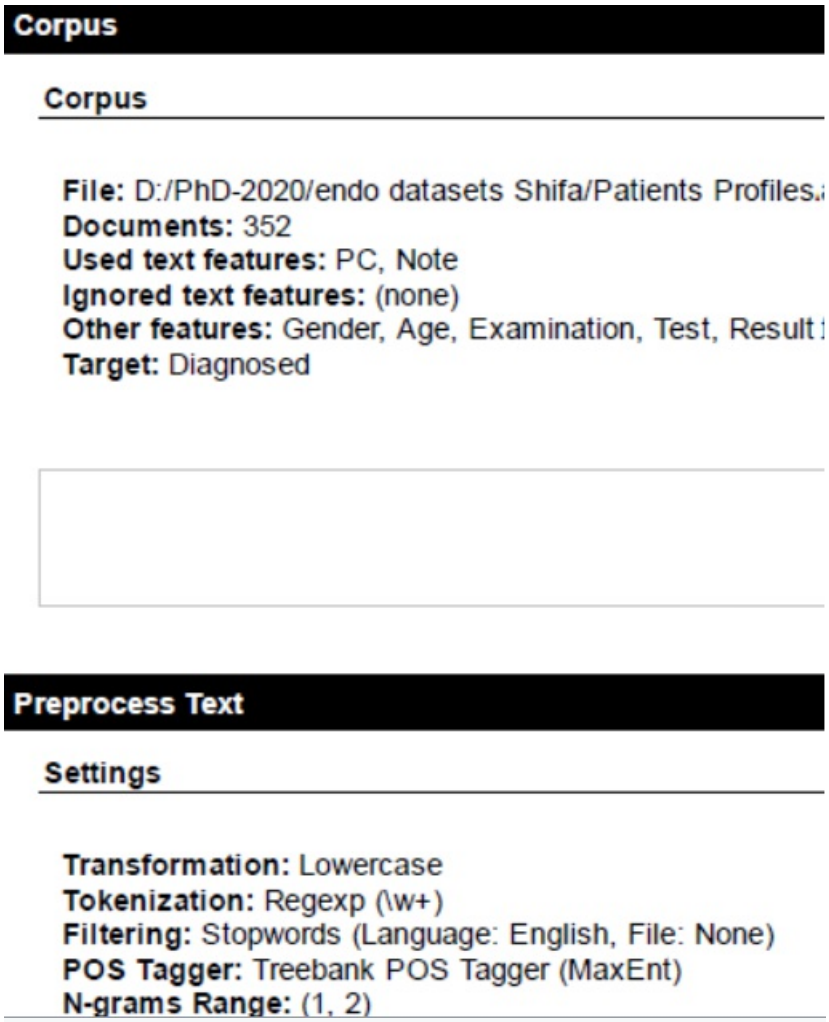



**Figure 46.** Male Patient of Age 73 with 352 records is converted to corpus to analyze text fields PC and Note.

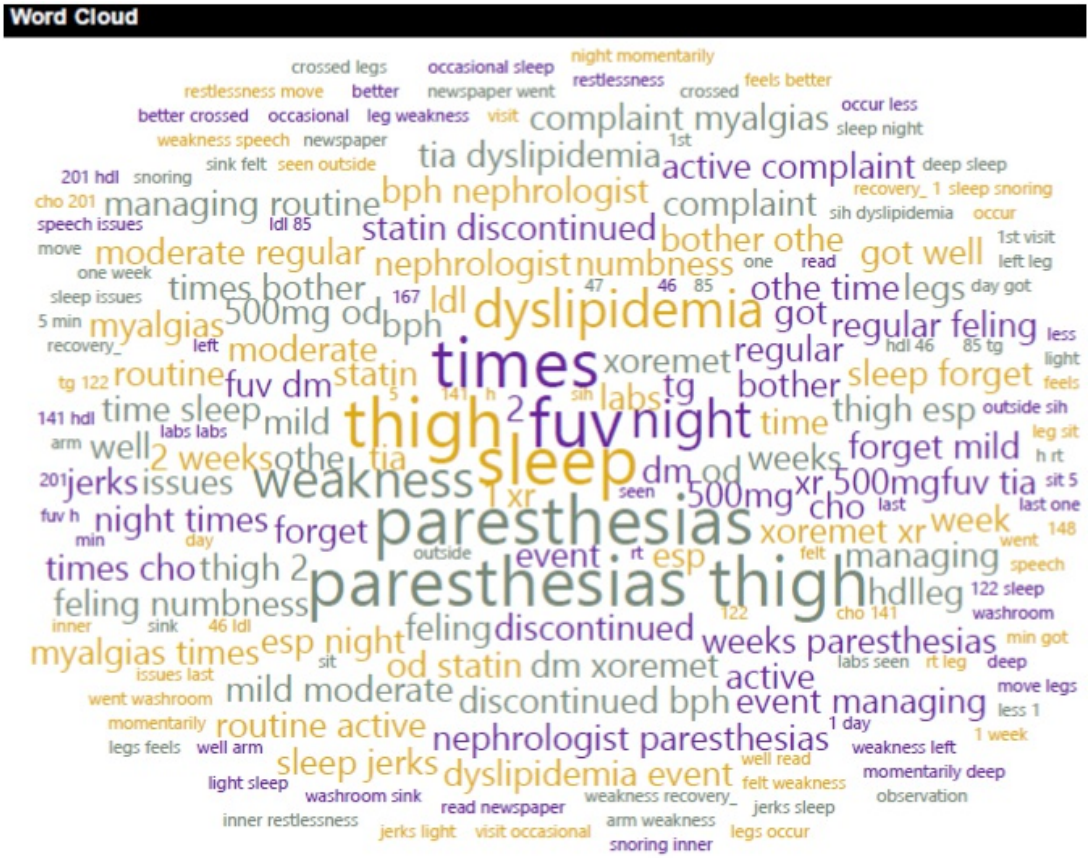



**Figure 47.** Word Cloud has the most weighed word 'sleep' in the patient 3 corpus.

It gave better prediction and test accuracy in terms of AUC, CA, F1-score, Precision and Recall when run in FAST-LMHFL analytics in Figure 47 compared to results of the LMHFL model applied on patient 3 in Figure 44.

Prediction accuracies for P3 in Set Pi = {AUC[0.958], CA[0.818], F1[0.818], Precision[0.833], Recall[0.818]}

Test accuracies for P3 in Set Pi = {AUC[0.965], CA[0.825], F1[0.825], Precision[0.834], Recall[0.825]}

Finally, the results are then seen in the confusion matrix as actual and predicted diagnoses.

**Confusion matrix for CN2 rule inducer (showing number of instances)**

| | | Predicted | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | BPH | DM | LUTS | Paresthesias | Sleep Starter (Jerk) | THORAT INFECTION | TIA | acne | ∑ |
| Actual | BPH | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | DM | 0 | 18 | 0 | 1 | 0 | 0 | 1 | 0 | 20 |
| | LUTS | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | Paresthesias | 0 | 1 | 0 | 18 | 1 | 0 | 0 | 0 | 20 |
| | Sleep Starter (Jerk) | 0 | 0 | 0 | 4 | 14 | 0 | 2 | 0 | 20 |
| | THORAT INFECTION | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | TIA | 0 | 1 | 0 | 2 | 1 | 0 | 16 | 0 | 20 |
| | acne | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | ∑ | | 20 | | 25 | 16 | | 19 | | 80 |

**Figure 48.** Confusion Matrix for male patient of age 73 with actual and predicted results for DM diagnosis and associated comorbidities using FAST-LMHFL analytics model showing the model's strength.

The other female patient with 3528 records referred to in Figure 45 before, suffered from chronic conditions with multiple diseases, mainly; breast cancer and DM were again analyzed using the Fast-LMHFL model. The limitations in text fields that were missing in most instances hindered the accurate and complete diagnosis of the patient as in Figure 49.

Likewise, there were several other patient profiles whose diagnosis was limited due to missing data fields.

**Confusion matrix for CN2 rule inducer (showing number of instances)**

| | | Predicted | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BERY BERY | CA BREAST | DM | DRUG REACTION | DYSPNEA | FRECKLES | JOINT PAIN | NO OBVIOUS CV CAUSE IDENTIFIED | PSORIASIS | S/P LAP CHOLE | pruritus | sore throat-cough | t pedis | ∑ |
| Actual | BERY BERY | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | CA BREAST | 0 | 18 | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 0 | 17 | 0 | 0 | 44 |
| | DM | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | DRUG REACTION | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | DYSPNEA | 0 | 1 | 0 | 0 | 14 | 0 | 0 | 10 | 0 | 0 | 19 | 0 | 0 | 44 |
| | FRECKLES | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | JOINT PAIN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | NO OBVIOUS CV CAUSE IDENTIFIED | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 31 | 0 | 0 | 13 | 0 | 0 | 44 |
| | PSORIASIS | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | S/P LAP CHOLE | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | pruritus | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 11 | 0 | 0 | 30 | 0 | 0 | 44 |
| | sore throat-cough | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | t pedis | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | ∑ | | 20 | | | 16 | | | 61 | | | 79 | | | 176 |

**Figure 49.** Female age 44 suffering from a chronic condition of DM with other underlying diseases could not be diagnosed completely with Fast-LMHFL due to missing text fields; Note and PC.

## 12. Combined Effect of All Features in Tabular and Text Datasets

Finally, the paper confirms the decision that DM is a chronic disease and is costly because of its widespread world population affecting different ethnic groups based on different geographical regions. The timely diagnosis of DM largely depends on diagnosing pre-diabetic patients through their family history and close monitoring of hemoglobin A1c (HbA1c) and glucose fasting (GF) levels to alert in case of spike and abnormality. Diagnosis of DM relates to other common

symptoms; obesity, blurred vision, aches and pains, hormonal, thyroid problems or during pregnancy, etc. WHO reported Saudi Arabia among the 10 most prominent countries for prevalence of DM [63]. Researchers believe that Pakistan statistics would not be far behind for ethnic similarity of people within both regions. Cost-effective diagnosis of DM was possible by selecting some common features. The paper [65], therefore, considered prominent screening and diagnostic features; HbA1c, fasting plasma glucose (FPG), High-density Lipoprotein Level (HDL), Low-density Lipoprotein Level (LDL), Hypertension (HTN), mobility level of patient, date of first diagnosis, primary and secondary diagnosis. The patient's phonotypical features selected were; date of birth, gender, height and weight, and location of patient. This dataset, thus, consisted of 16 features from 3000 anonymous patients admitted from 2016 to 2018, in different departments; inpatient, outpatient and emergency from five hospitals in Saudi Arabia residing in the central region, eastern region and western region. This dataset was preprocessed to clean and prune missing values. Patients below the age of 19 or females having gestational diabetes were removed to keep T2D in focus. The instances that remained were only 162. The primary and secondary diagnosis features were removed to keep the diagnosis label diabetic or non-diabetic only. The date of birth (DOB) feature was replaced with Age. The region and start date of diagnosis were also removed for classification accuracy. The remaining features were only 10; age, gender, height, weight, HTN, Physical Activity Level (PAL), Lipoprotein levels (HDL and LDL), FPG and HbA1c. This final dataset had to run through five ML classifiers in the scikit-learn library, therefore, it was converted to numeric for efficient processing. This dataset was converted to two datasets where one dataset had an FPG label to compare with results from HbA1c in other datasets with common features in both datasets.

In [65] features are selected based on their importance through scores given for the relevance in target labeling. There are many ways in which features can be ranked either through coefficient values in SVM or logistic regression, or split metrics decision trees or random forest, and correlation between features also helps in finding the relevance and significance. However, basic algorithms like; SVM, logistic regression, decision trees or random forests have become old concepts for finding linear or scalar relationships in small datasets. Therefore, techniques like recursive feature elimination (RFE) are now preferred to know the impact of an exploratory feature on a response variable. Another technique is permutation, where feature importance is observed through model accuracy on random shuffling. Lastly, feature elimination lets you truncate the redundant input nodes that are non-informative for the best prediction model. These methods for best feature selection not only reduce the cost of computation but also increase model performance by making it more uniform and generalized to adapt. Spearman and Pearson correlation metrics in addition to the permutation technique are found effective for the selection of relevant features. Application of distance metrics within the predictive models helps cluster all the nodes as communities and sub-communities that are best depicted in tree structure through hierarchical clustering for clarity of resultant associations and relationships between features and the target labels as in our study; diagnosis and ICD-10-CM. In [63], the classifiers with the same configurations were run on the two datasets in two steps with nine features and with eight features by eliminating Gender = M in the final step. Performance evaluation was done using the widely accepted parameters; accuracy, recall, precision and F1-score with 10-fold cross-validation. As a result, FPG gave better predictions than HbA1c for all five classifiers. SVM performed better on the HbA1c dataset and no change in performance was observed with feature reduction while other classifiers showed increased or decreased performance. Decision Tree (DT) and Ensemble Majority Voting (EMV) performed better with reduced features

while random forest (RF) performance decreased with reduced features [63].

On analysis [63], the correlation between HbA1c and FPG with other features like; LDL, PAL, HTN, Age, height and weight, also depicted very significant results. LDL or bad cholesterol is more related to HbA1c than FPG. HTN, age and weight are correlated while inversely proportional to PAL. HTN also correlates to increased HbA1c and FPG with a similar impact. Height has a negative correlation with HbA1c and FPG which shows that shorter people are at more risk of diabetes. The weight of a person has no correlation with diabetes as it is observed that females are more prone to diabetes but have less weight than men in general. HDL gets better with increased PAL. PAL has no direct correlation with HbA1c or FPG but does have an effect on other features.

Researchers doing the study in [63] came up with very valuable suggestions for adding up other important vitals like; blood pressure, temperature, waist size or vision, etc. to monitor the prescreening of diabetic patients. These researchers also felt limited in access to a large dataset as this dataset was smaller even with 3000 patients and had only 162 records after the elimination of records with missing values. A larger dataset was needed that promised better insights and greater performance in analytics as this paper explained through multiple experiments and results. In this paper, further going deep into the features list on expansion, as in Figure 50, we carefully articulated the relations depicted by the resulting inferences and have elaborated in this section.

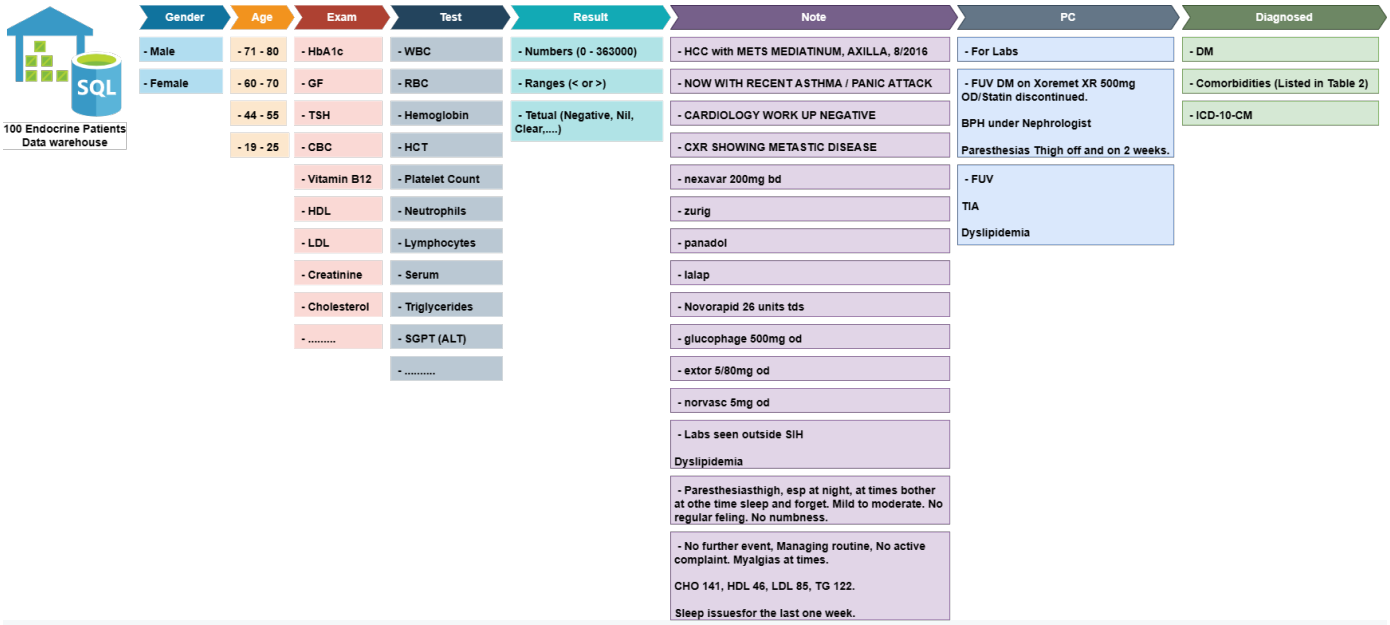



**Figure 50.** 9304 pruned and cleaned records of 100 patients in our Data warehouse and as batches of individual patient profiles gave us very informative insights and associations between a set of 7 features for the diagnosis and prognosis of DM and its various comorbidity diseases.

Heat Maps in Figures 17 and 18 of section 7 depict a very good picture of different features related to or resulting in different diagnoses; higher HbA1c, LDL, GFR, or abnormal urine calcium results in DM. Then in Figure 47, we see textual features popping out are; sleep, weakness, forget, jerks, and numbness and then there are some medicine names like; statin and xoremet, etc. with diseases as; myalgia, dyslipidemia or nephrologist, all signifying associations to DM and

other common diseases that co-occur with it. Figure 50 also elaborates how the considered features for rule sets look when expanded in combination with neural nets in deep multinomial or graph clusters as in LMHFL resulting in the diagnosis and prognosis of DM and its comorbidities.

The auto feature evaluation is seen in Figure 46. The same features are shown in a descriptive manner in Figure 50 and later are expanded for detailed elaboration in Tables 3 and 4. Table 3 shows all 109 tests residing in the dataset that went through simulated analytics in open-source cloud platforms with all other features taken into account to diagnose. Table 4 elaborates in detail the diagnostic counts for different age groups in relation to DM and its co-occurring diseases with respect to individual patients specified as Male/Female with the selection of test ranges and the results. It is clear that here the stats have been taken into account and focused on HbA1c, HDL, LDL, GF, TSH and pH tests to know their significance. In Table 4 we also see the undiagnosed disease records labeled as ‘nan’ and notice that in some diagnoses either HbA1c or TSH were enough to diagnose a patient with DM or its comorbidity disease.

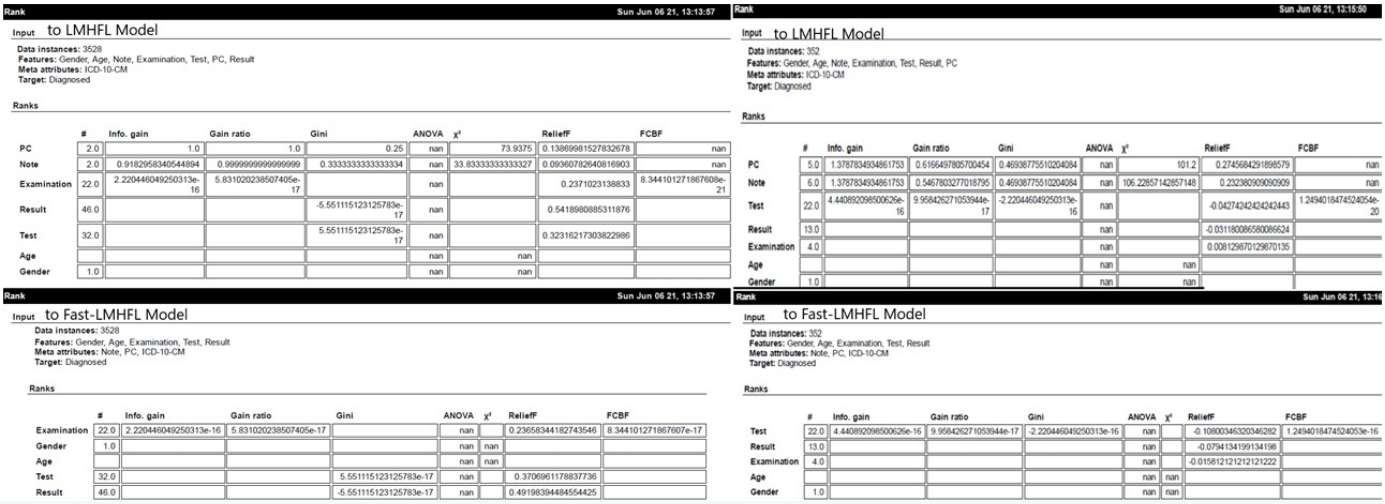



**Figure 51.** Auto Features Evaluation Matrix generated on two individual patients' datasets of 3528 and 352 records that were input in LMHFL/Fast-LMHFL Models in Orange Framework.

**Table 4.** The complete list of tests in our dataset 2 having 100 patients with 15696 records.

| Test | TestID | Test | TestID | Test | TestID |
|---|---|---|---|---|---|
| **24 Hr Urine Calcium** | 0 | Gamma G.T. | 37 | Phosphorous | 74 |
| **24 Hr Urine Cortisol** | 1 | Glucose | 38 | Platelet Count | 75 |
| **24 Hr Urine volume** | 2 | Glucose Fasting | 39 | Potassium Serum | 76 |
| **25-Hydroxy Vitamin D** | 3 | Glucose P.P 2 Hour | 40 | Prolactin | 77 |
| **ACR (Spot Urine Albumin/Creatinine ratio)** | 4 | Glucose Random | 41 | Protein | 78 |
| **ACTH** | 5 | Gram Negative Bacilli | 42 | RBC | 79 |
| **Albumin Serum** | 6 | Gram Negative Cocci | 43 | RBC's | 80 |
| **Alkaline Phosphatase** | 7 | Gram Positive Bacilli | 44 | RBC, Total | 81 |
| **Anti - Thyroglobulin** | 8 | Gram Positive Cocci | 45 | RDW | 82 |
| **Anti - Thyroid Peroxidase** | 9 | Gram Stain | 46 | Red Blood Cells | 83 |
| **Appearance** | 10 | Gross Blood | 47 | SGOT(AST) | 84 |
| **BUN** | 11 | HCT | 48 | SGPT (ALT) | 85 |
| **Bacteria** | 12 | HbA1c | 49 | Sodium Serum | 86 |
| **Basophil** | 13 | Hemoglobin | 50 | Source | 87 |
| **Bicarbonate Serum** | 14 | High Density Lipoprotein(HDL) | 51 | Specific Gravity | 88 |
| **Bilirubin** | 15 | Insulin | 52 | Squamous Epithelial Cells | 89 |
| **Blood** | 16 | Intact PTH | 53 | T3 | 90 |
| **Calcium** | 17 | Ionized Calcium | 54 | T4 | 91 |
| **Cast** | 18 | Iron Serum | 55 | TSH | 92 |
| **Chloride Serum** | 19 | Ketone | 56 | Testosterone | 93 |
| **Cholesterol** | 20 | LH | 57 | Total Bilirubin | 94 |
| **Clostridium difficile antigen** | 21 | Leukocytes-Estrases | 58 | Triglycerides | 95 |
| **Color** | 22 | Low Density Lipoprotein (LDL) | 59 | Urea Serum | 96 |
| **Consistency** | 23 | Lymphocytes | 60 | Uric Acid | 97 |
| **Cortisol** | 24 | MCH | 61 | Urine Calcium | 98 |
| **Creatinine** | 25 | MCHC | 62 | Urine Cortisol | 99 |
| **Crystals** | 26 | MCV | 63 | Urine Micro Albumin | 100 |
| **Culture** | 27 | MPV | 64 | Urobilinogen | 101 |
| **Cyst** | 28 | Monocytes | 65 | Vitamin B12 | 102 |
| **Direct Bilirubin** | 29 | Mucus | 66 | WBC | 103 |
| **ESR (Wester-gren)** | 30 | Neutrophils | 67 | WBC Total | 104 |
| **Eosinophils** | 31 | Nitrite | 68 | WBC's | 105 |
| **Epithelial Cells** | 32 | OH Progesterone | 69 | Worm | 106 |
| **Estimated GFR using CKD-EPI (PK)** | 33 | Other | 70 | Yeast | 107 |
| **FREE T3** | 34 | Ova | 71 | pH | 108 |
| **FSH** | 35 | POCT Glucose | 72 | | |
| **Free-T4** | 36 | PUS Cells | 73 | | |

**Table 5.** Some stats of Diseases co-occurring with DM related to relevant Tests and its Results for different age groups in dataset 2 of 100 patients categorized as Male/Female.

| | | | | | | | | | —Age | | —Result | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gender | PatientID | Disease | HbA1c | GF | HDL | LDL | TSH | pH | count | Mean | count | Mean |
| | | DM | 5.7% ¡ HBA 1c ¡ 6.4% | nan | nan | nan | nan | nan | 3 | 49 | 3 | 6.3 |
| | | | HbA1c ¿= 6.5% | nan | nan | nan | nan | nan | 9 | 49 | 9 | 8.03 |
| | 132303 | | nan | nan | HDL ¡ 50 | nan | nan | nan | 15 | 49 | 15 | 33.6 |
| | | | | | nan | LDL ¿= 100 | nan | nan | 15 | 49 | 15 | 162.2 |
| | | | | | | nan | nan | nan | 66 | 49 | 57 | 140.3 |
| Female | | Low back pain | 5.7% ¡ HBA 1c ¡ 6.4% | nan | nan | nan | nan | nan | 1 | 49 | 1 | 6.3 |
| | | | HbA1c ¿= 6.5% | nan | nan | nan | nan | nan | 3 | 49 | 3 | 8.03 |
| | | | nan | nan | HDL ¡ 50 | nan | nan | nan | 5 | 49 | 5 | 33.6 |
| | | | | | nan | LDL ¿= 100 | nan | nan | 5 | 49 | 5 | 162.2 |
| | | | | | | nan | nan | nan | 22 | 49 | 19 | 140.3 |
| | | nan | 5.7% ¡ HBA 1c ¡ 6.4% | nan | nan | nan | nan | nan | 3 | 49 | 3 | 6.3 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | HbA1c ¿= 6.5% | nan | nan | nan | nan | nan | 9 | 49 | 9 | 8.03 |

## 13. Analysis with Explained Results for Future Recommendation

This paper has evaluated the functionalities of different cloud platforms on different sizes of datasets with automated ML model simulations with varying accuracy results on different hyperparameter settings on patients' phenotypically [50] defined data for diagnosis as in Table 1. Customized rules are extracted for the diagnosis of diabetes mellitus and its comorbidities identified in various datasets provided. Most relevant features in the endocrine datasets for multiple visits of patients are now known. Datasets are purified and pruned (Table 2) for faster processing. Exhaustive experimentation with advanced ML and NLP tools on several open-source cloud platforms led to more accurate diagnostic results up to 100% as achieved with Deep Multinomial Learning with different settings for multiclass or multitask classification in the future as summarized in our recent conference and research papers [64][65].

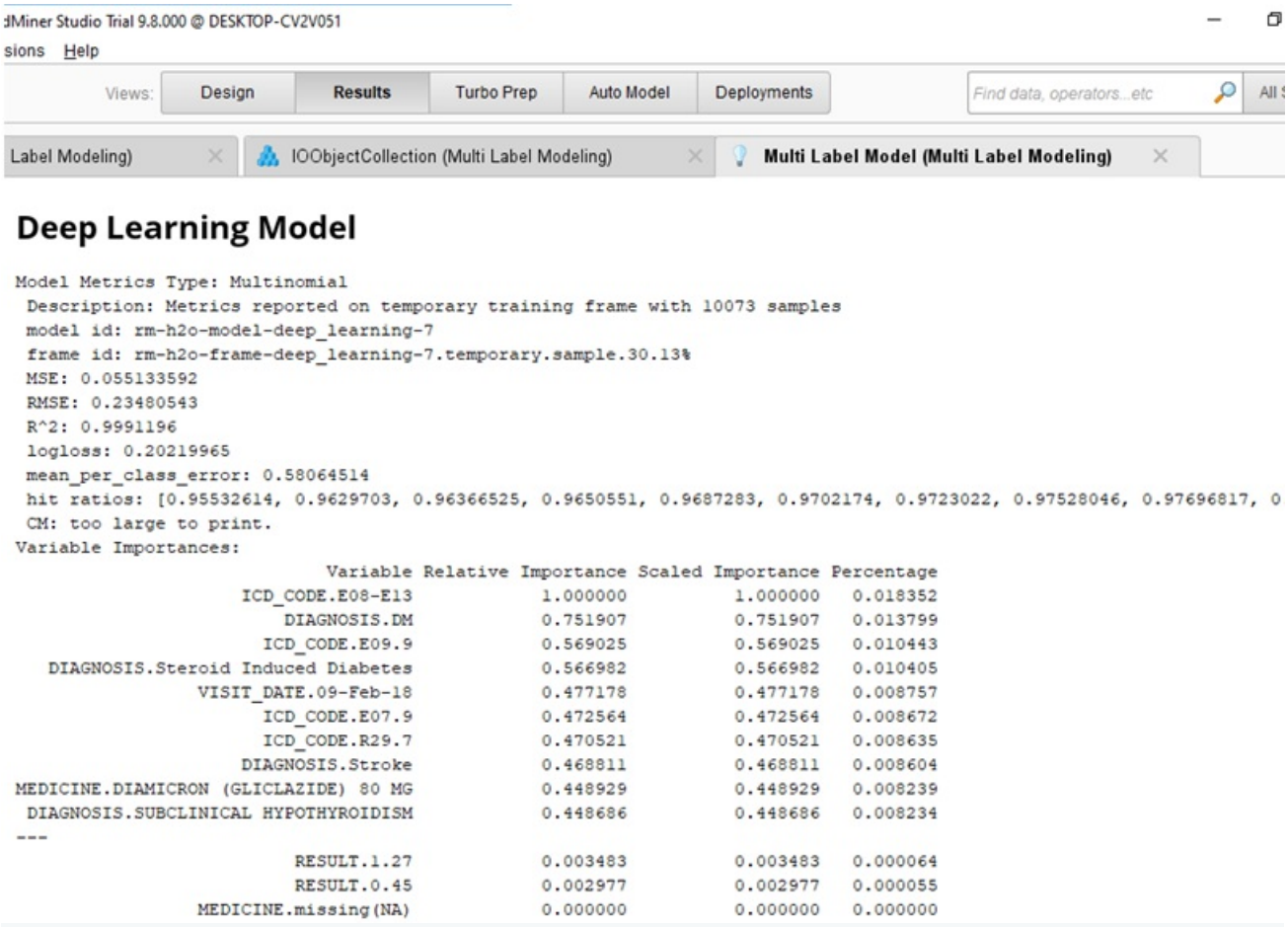



**Figure 52.** Descriptive accuracy results with log loss equaling 0.202 on 2 hidden layers of sizes 31 and 36.

The pruned big data set of 87803 records was reduced to 33,185 and was found limited for single patient profile diagnosis class which became constant as the patient was either diagnosed with DM or its comorbidity disease. This was not the

case with 100 patients' records of 3650 and 15696 instances as individual patients had DM and related comorbidity disease as listed in Table 2. The datasets from 100 patients gave us induced rule sets through LMHFL designed and simulated on Orange Framework (Figure 35). Diseases are given standard ICD-10-CM codes for analysis on the complete diabetic patient diagnostic profile that led to the diagnosis of comorbidities formed later with time.

Figure 6, 8 and 9 are good representations formed using LMHFL run on 3650 records for individual diagnosis of patients, of Dataset 1, based on their exam and tests results. Figure 34 also shows each patient diagnosed with DM and its comorbidities, a representation from 15696 records from Dataset 2. In a detailed analysis of these 100 patients after pruning only three individual patient profiles were fetched for having DM and its comorbidity diseases with concluding results depicted in Figure 44 as a confusion matrix;

a. Patient 1: Male, Age: 53, 24 records, DM diagnosed having two comorbidity diseases; Benign Prostatic Hyperplasia (BPH), and Premature Ejaculation (PME)
b. Patient 2: Female, Age: 49, 144 records, Diagnosed; DM with Comorbidity: Lower Back Pain (LBP)
c. Patient 3: Male, Age: 73, 352 records, Diagnosed DM and Comorbidities; Transient Ischemic Attack (TIA), BPH, Lower Urinary Tract Symptoms (LUTS), acne, sleep starter (Jerk), Paresthesias and Throat Infection.

The representations visualized through analytics assisted us in finding hidden patterns for extracting custom rule sets outlined in section 4. Further, the complete sets of analytics are applied to textual data without transforming it into fuzzy data.

15696 instances from 100 patients gave individual patient representations for DM and its comorbidities using auto Deep Multinomial Learning analytics over RapidMiner platform as in Figure 22 and 23. Finally, Deep Multinomial Learning with Multi-Label operator showed good speed ranging from 23s (Figure 25 and 27) to 3mins (Figure 30) to 6mins (Figure 36) depending on the size of the dataset provided reached to a maximum of 100% accuracy in multiple runs. The datasets considered in experiments were tested raw and later in cleaner and pruned form for higher speed and accuracy. The datasets gradually grew in size to be processed on a uniform data model to test generalization on standard FHIR v4.0 HL7 Schema with the flexibility to add or drop features based on desired results.

We further tried LMHFL in combination with Fastai for textual tabular data processing and analysis results got even better and near to 100% accuracy as may be visualized in the confusion matrix in Figure 48. The errors depicted are due to the strong association between DM and its comorbidities and the similarity in the feature set used to diagnose multiple diseases in a single patient. A female patient with breast cancer and DM was also carefully analyzed for the limitation posed in diagnosis due to missing data and the close association of DM and associated diseases with her breast cancer.

In the future, researchers are interested in tuning, testing and validating LMHFL hybridized with other deep learning analytics models like MLR and multi-label ensemble models for multi-label classification with up to 100% accuracy and fair distinguishable diagnostic results through EHR big data removing any chance of algorithmic biasness. There are still some forms of diabetes and its comorbidities highlighted in [22][66] like; Hepatogenous Diabetes (HD) and Antecedent Diabetes (AD) leading to Liver Failure or liver cirrhosis respectively missing in our datasets with other features, shown in

Table 2, and a recent occurrence of Covid-19, that the researchers intend to include in future. Researchers intend to make the proposed models more generalized having a dynamic training dataset being continuously updated by feedback and revised diagnostic results from healthcare practitioners for diagnosis with maximum accuracy [66][67].

## Declarations

### Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Authors' Contributions

This is an original research emerged with the collaborated efforts of the PhD(CS) Scholar and her professors.

- Sarah Shafqat conceptualized this study, designed the unified clinical data model, extracted, processed and prepared the corpora for exhaustive experimentation and analysis.
- Prof Dr Zahid reviewed, proofread and structured the paper with the scholar. He gave suggestions where necessary and assisted the scholar to draw Figure 1 of high-level diagnostic architecture that forms the basis of her research.
- Prof Dr. Raihan Rasool is an employee of IBM Technologies, and the views are of his own.
- Prof Dr Qaisar managed the collaboration from the platform of International Islamic University Islamabad (IIUI), Pakistan.
- Prof. Dr Hafiz Farooq Ahmad directed and led this research with the approval of International Islamic University Islamabad (IIUI), Pakistan.

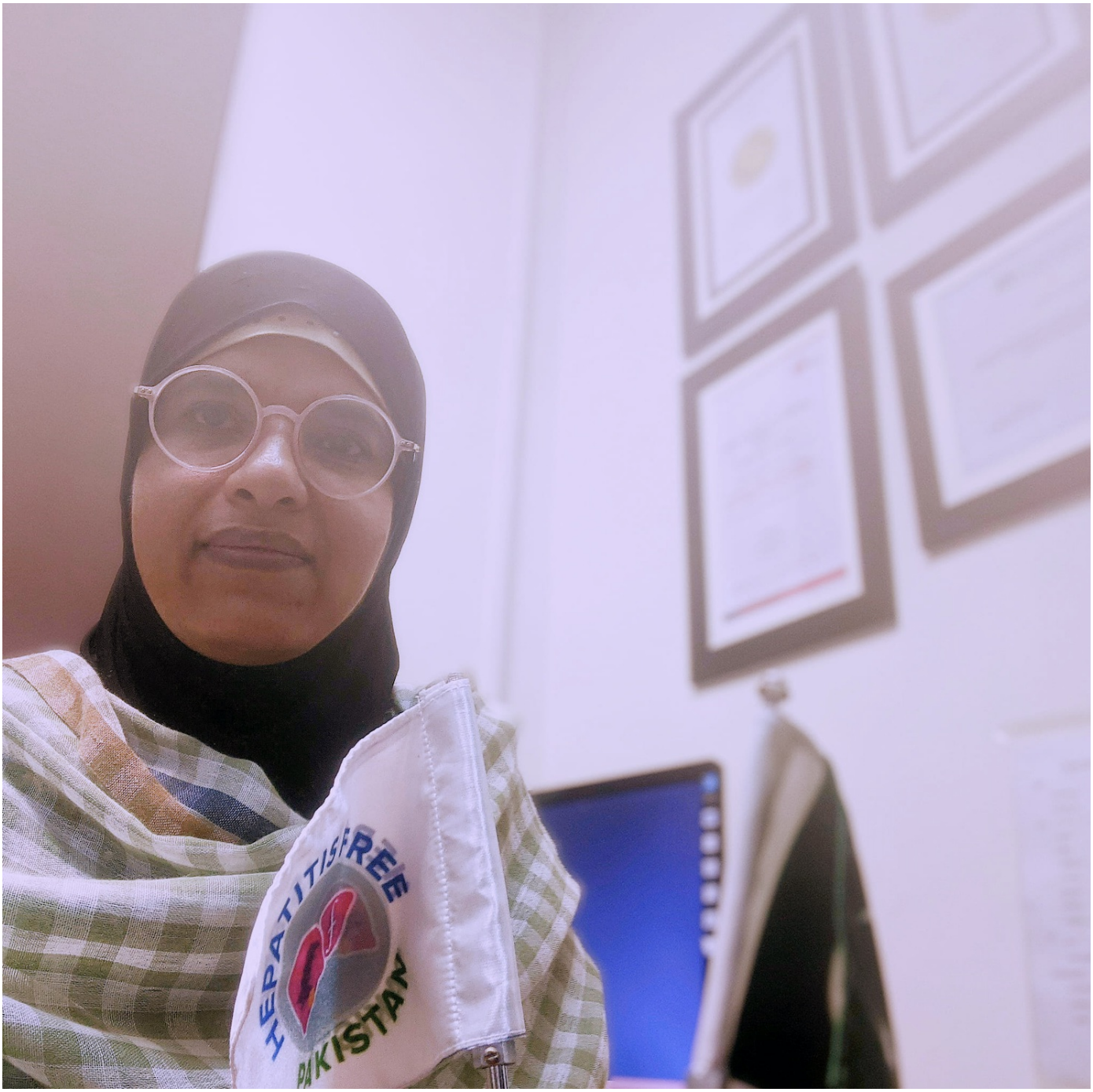
### Data Confidentiality Statement

Due to the sensitive nature of the real-time patients' EHR data analyzed in this study, hospital and other executive stakeholders are assured that it would remain confidential and would not be shared.

### Availability of Data

Sarah Shafqat, extracted endocrine EHR big data under Grant (IRB\# 996-271-2018) approved by hospital’s MIS department. This dataset was further processed during the research for analysis and would become a unified knowledge base. Scholar plans to apply for patent for the analytical models and the medical corpora for further research and collaboration.

## About the Authors




**Sarah Shafqat.** She is a profound research scholar in field of Information Technology having sound knowledge of Business Processes. She received Bachelor of Business (Information Technology) from Curtin University, Australia and continued her career with Master's degree in Business Administration with majors in Human Resource Management and post-graduation in Software Engineering & Computer Science. She is fully aware of the underlying risks in organizations residing over sensitive data like healthcare and individuals commuting over the cloud for in-time services. In spite of all the risks that are associated with the cloud she fully understands its importance and potential growth in the services industry. Therefore, the implementation of security in cloud infrastructure is her keen interest. Currently, she has some

highly resourceful publications in renowned international journals. Her research area revolves around cloud computing, big data analytics and healthcare informatics.

**Raihan Ur Rasool.** Raihan has an accomplished tech career that spans over 23 years with success stories around product development, design and operations of large distributed systems, security and transformations, and scientific research. With a doctorate in distributed systems and having published a book and over 80 international papers, Raihan has been at the forefront of innovations. He is currently with IBM and looks after the Automation portfolio as a senior architect.

_____

*This is an original research emerged with the collaborated efforts of the PhD(CS) Scholar and her professors. The views given here are by the authors themselves.*

*This research is directed and led by Prof. Dr Hafiz Farooq Ahmad with the approval of International Islamic University Islamabad (IIUI), Pakistan.*

## Other References

- *W. L. Lowe et al., "Hyperglycemia and Adverse Pregnancy Outcome Follow-up Study (HAPO FUS): Maternal Gestational Diabetes Mellitus and Childhood Glucose Metabolism.," Diabetes Care, vol. 42, no. 3, pp. 372–380, Mar. 2019, doi: 10.2337/dc18-1646.*
- *J. Luo, L. Phillips, S. Liu, J. Wactawski-Wende, and K. L. Margolis, "Diabetes, diabetes treatment, and risk of thyroid cancer," J. Clin. Endocrinol. Metab., vol. 101, no. 3, pp. 1243–1248, 2016, doi: 10.1210/jc.2015-3901.*

## References


1. [a, b] *S. J. Appel, T. M. Wadas, R. S. Rosenthal, and F. Ovalle, "Latent autoimmune diabetes of adulthood (LADA): An often misdiagnosed type of diabetes mellitus," J. Am. Acad. Nurse Pract., vol. 21, no. 3, pp. 156–159, 2009, doi: 10.1111/j.1745-7599.2009.00399.x.*
2. [^] *D. S. Gardner and E. S. Tai, "Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy Clinical features and treatment of maturity-onset diabetes of the young (MODY)," Diabetes, Metab. Syndr. Obes. Targets Ther., vol. 5, pp. 101–108, 2012, doi: 10.2147/DMSO.S23353.*
3. [^] *A. S. Shah et al., "the SEARCH for Diabetes in Youth Study," vol. 25, no. 0, pp. 717–721, 2015, doi: 10.1515/jpem-2012-0070.Adiponectin.*
4. [a, b, c, d, e, f] *I. Technology and C. Members, "Investigating Multi-layer Machine Learning Algorithms to Improve Diabetic Analytic Models Investigating Multi-layer Machine Learning Algorithms to Improve Diabetic Analytic Models," no. April, 2018.*
5. [a, b, c, d] *S. A. D. Alalwan, "Diabetic analytics: Proposed conceptual data mining approaches in type 2 diabetes dataset,"*

*Indones. J. Electr. Eng. Comput. Sci., vol. 14, no. 1, pp. 92–99, 2019, doi: 10.11591/ijeecs.v14.i1.pp92-99.*

6. ^*H. Qiu et al., "Electronic Health Record Driven Prediction for Gestational Diabetes Mellitus in Early Pregnancy," Sci. Rep., vol. 7, no. 1, p. 16417, 2017, doi: 10.1038/s41598-017-16665-y.*
7. a, b*D. De Silva, F. Burstein, H. Jelinek, and A. Stranieri, "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, doi: 10.3127/ajis.v19i0.1183.*
8. ^*N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," J. Big Data, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0175-6.*
9. a, b*M. Alehegn, R. Joshi, and P. Mulay, "Analysis and prediction of diabetes mellitus using machine learning algorithm," Int. J. Pure Appl. Math., vol. 118, no. Special Issue 9, 2018.*
10. a, b*S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: management, analysis and future prospects," J. Big Data, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0217-0.*
11. a, b*A. Srinivasan, "Essays on Digital Health and Preventive Care Analytics Item Type text; Electronic Dissertation," 2019. Accessed: Nov. 26, 2019. [Online]. Available: http://hdl.handle.net/10150/632971.*
12. a, b*A. Wang, "A Deep Learning Based Transfer Learning Framework for Healthcare Text Analytics," 2019.*
13. a, b*A. Zamuda, C. Zarges, G. Stiglic, and G. Hrovat, "Stability selection using a genetic algorithm and logistic linear regression on healthcare records," 2017, pp. 143–144, doi: 10.1145/3067695.3076077.*
14. a, b, c*J. Waring, C. Lindvall, and R. Umeton, "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare," Artificial Intelligence in Medicine, vol. 104. Elsevier B.V., Apr. 01, 2020, doi: 10.1016/j.artmed.2020.101822.*
15. a, b, c, d*B. Liu, Y. Li, S. Ghosh, Z. Sun, K. Ng, and J. Hu, "Complication Risk Profiling in Diabetes Care: A Bayesian Multi-Task and Feature Relationship Learning Approach," IEEE Trans. Knowl. Data Eng., vol. XX, 2019, doi: 10.1109/TKDE.2019.2904060.*
16. ^*A. Talaei-Khoei, M. Tavana, and J. M. Wilson, "A predictive analytics framework for identifying patients at risk of developing multiple medical complications caused by chronic diseases," Elsevier, 2019, doi: 10.1016/j.artmed.2019.101750.*
17. ^*D. G. Benjamin Oberman, Aliasgher Khaku, Fabian Camacho, "Relationship between obesity, diabetes and the risk of thyroid cancer," Am. J. Otolaryngol., vol. 36, no. 04, pp. 535–541, 2015, doi: https://doi.org/10.1016/j.amjoto.2015.02.015.*
18. ^*D. O. F. Diabetes, "Diagnosis and classification of diabetes mellitus," Diabetes Care, vol. 33, no. SUPPL. 1, 2010, doi: 10.2337/dc10-S062.*
19. ^*N. S. Kakoly, A. Earnest, H. J. Teede, L. J. Moran, and A. E. Joham, "The impact of obesity on the incidence of type 2 diabetes among women with polycystic ovary syndrome," Diabetes Care, vol. 42, no. 4, pp. 560–567, Apr. 2019, doi: 10.2337/dc18-1738.*
20. a, b, c*T. M. Ramachandran, A. H. R. Rajneesh, G. S. Zacharia, and R. P. Adarsh, "Cirrhosis of liver and diabetes mellitus: The diabolic duo?," J. Clin. Diagnostic Res., vol. 11, no. 9, pp. OC01–OC05, 2017, doi: 10.7860/JCDR/2017/30705.10529.*

21. [a, b]*P. Klimek, A. Kautzky-Willer, A. Chmiel, I. Schiller-Frühwirth, and S. Thurner, "Quantification of Diabetes Comorbidity Risks across Life Using Nation-Wide Big Claims Data," PLoS Comput. Biol., vol. 11, no. 4, pp. 1–16, 2015, doi: 10.1371/journal.pcbi.1004125.*
22. [a, b]*L. Porepa, J. Ray, P. Sanchez-Romeu, G. B.- Cmaj, and undefined 2010, "Newly diagnosed diabetes mellitus as a risk factor for serious liver disease," Can Med Assoc, 2010, doi: 10.1503/cmaj.092144.*
23. [a, b, c]*P. Dworzynski et al., "Nationwide prediction of type 2 diabetes comorbidities," bioRxiv, p. 664722, 2019, doi: 10.1101/664722.*
24. [a, b]*D. D. M. Agany, J. E. Pietri, and E. Z. Gnimpieba, "Assessment of vector-host-pathogen relationships using data mining and machine learning," Computational and Structural Biotechnology Journal, vol. 18. Elsevier B.V., pp. 1704–1721, Jan. 01, 2020, doi: 10.1016/j.csbj.2020.06.031.*
25. [^]*L. Subirats, R. Gil, and R. García, "Personalization of ontologies visualization: Use case of diabetes," in Studies in Computational Intelligence, vol. 815, 2019, pp. 3–24.*
26. [a, b, c]*A. Z. Woldaregay et al., "Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes," Artificial Intelligence in Medicine, vol. 98. Elsevier B.V., pp. 109–134, Jul. 01, 2019, doi: 10.1016/j.artmed.2019.07.007.*
27. [^]*J. S. Sartakhti, M. H. Zangooei, and K. Mozafari, "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)," Comput. Methods Programs Biomed., vol. 108, no. 2, pp. 570–579, 2012, doi: 10.1016/j.cmpb.2011.08.003.*
28. [a, b]*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.*
29. [a, b]*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.*
30. [a, b]*M. Shuja, S. Mittal, and M. Zaman, "Diabetes Mellitus and Data Mining Techniques A survey," Int. J. Comput. Sci. Eng., vol. 7, no. 1, pp. 858–861, 2019, doi: 10.26438/ijcse/v7i1.858861.*
31. [a, b]*I. Dankwa-Mullan, M. Rivo, M. Sepulveda, Y. Park, J. Snowdon, and K. Rhee, "Transforming Diabetes Care Through Artificial Intelligence: The Future Is Here," Popul. Health Manag., vol. 22, no. 3, pp. 229–242, Jun. 2019, doi: 10.1089/pop.2018.0129.*
32. [a, b]*G. Fico et al., "What do healthcare professionals need to turn risk models for type 2 diabetes into usable computerized clinical decision support systems? Lessons learned from the MOSAIC project," BMC Med. Inform. Decis. Mak., vol. 19, no. 1, Aug. 2019, doi: 10.1186/s12911-019-0887-8.*
33. [a, b, c, d, e, f]*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.*
34. [a, b]*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.*
35. [^]*M. Raghu and E. Schmidt, "A Survey of Deep Learning for Scientific Discovery," arXiv. pp. 1–48, 2020.*
36. [^]*K. Ng, J. Sun, J. Hu, and F. Wang, "Personalized Predictive Modeling and Risk Factor Identification using Patient*

*Similarity.," AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci., vol. 2015, pp. 132–6, 2015, Accessed: Dec. 18, 2020. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26306255.*

37. ^*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.*
38. a, b, c, d, e, f *C. Lam, D. Yi, M. Guo, T. L.-A. summits on translational, and undefined 2018, "Automated detection of diabetic retinopathy using deep learning," ncbi.nlm.nih.gov, Accessed: Nov. 30, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/pmc5961805/.*
39. a, b, c, d, e *J. Luo, C. M. Wong, and C. M. Vong, "Multinomial Bayesian extreme learning machine for sparse and accurate classification model," Neurocomputing, vol. 423, pp. 24–33, Jan. 2021, doi: 10.1016/j.neucom.2020.09.061.*
40. ^*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.*
41. ^*S. Piri, D. Delen, T. Liu, and W. Paiva, "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, doi: 10.1016/j.eswa.2017.09.061.*
42. ^*M. Vamvini, V.-A. Lioutas, and R. J. W. Middelbeek, "Characteristics and Diabetes Control in Adults With Type 1 Diabetes Admitted With COVID-19 Infection," Diabetes Care, vol. 43, no. October, p. dc201540, 2020, doi: 10.2337/dc20-1540.*
43. ^*V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," J. Stat. Mech. Theory Exp., vol. 2008, no. 10, pp. 1–12, 2008, doi: 10.1088/1742-5468/2008/10/P10008.*
44. ^*M. Saint-Guillain, "Automatic detection of community structures in networks," pp. 1–13, 2012.*
45. ^*V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," Sci. Rep., vol. 9, no. 1, pp. 1–12, Dec. 2019, doi: 10.1038/s41598-019-41695-z.*
46. ^*J. Howard and S. Gugger, "Fastai: A layered api for deep learning," Inf., vol. 11, no. 2, 2020, doi: 10.3390/info11020108.*
47. ^*J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 2018, vol. 1, pp. 328–339, doi: 10.18653/v1/p18-1031.*
48. a, b, c, d, e *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.*
49. a, b *M. Savić, V. Kurbalija, Z. Bosnić, and M. Ivanović, "Feature selection based on community detection in feature correlation networks," Computing, vol. 101, no. 10, pp. 1513–1538, 2019, doi: 10.1007/s00607-019-00705-8.*
50. a, b *Y. Halpern, S. Horng, Y. Choi, and D. Sontag, "Electronic medical record phenotyping using the anchor and learn framework," J. Am. Med. Informatics Assoc., vol. 23, no. 4, pp. 731–740, 2016, doi: 10.1093/jamia/ocw011.*
51. ^*A. Palvanov and Y. I. Cho, "Comparisons of Deep Learning Algorithms for MNIST in Real-Time Environment," vol. 18, no. 2, pp. 126–134, 2018.*
52. ^*H. M. Proença and M. van Leeuwen, "Interpretable multiclass classification by MDL-based rule lists," Inf. Sci. (Ny).,*

*vol. 512, pp. 1372–1393, 2020, doi: 10.1016/j.ins.2019.10.050.*

53. [a, b]*P.-W. Wang, P. L. Donti, B. Wilder, and Z. Kolter, "SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver," ieeexplore.ieee.org, 2019, Accessed: May 04, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8086133/.*
54. [^]*P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, "Deepr: A Convolutional Net for Medical Records," pp. 1–9, 2016, [Online]. Available: http://arxiv.org/abs/1607.07519.*
55. [^]*B. Wang and I. Davidson, "Towards Fair Deep Clustering With Multi-State Protected Variables," Jan. 2019, Accessed: Feb. 25, 2019. [Online]. Available: http://arxiv.org/abs/1901.10053.*
56. [^]*J. Zhang, Y. Wang, P. Molino, L. Li, and D. S. Ebert, "Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models," IEEE Trans. Vis. Comput. Graph., vol. 25, no. 1, pp. 364–373, 2019, doi: 10.1109/TVCG.2018.2864499.*
57. [^]*D. Kobak, G. Linderman, S. Steinerberger, Y. Kluger, and P. Berens, "Heavy-Tailed Kernels Reveal a Finer Cluster Structure in t-SNE Visualisations," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020, vol. 11906 LNAI, pp. 124–139, doi: 10.1007/978-3-030-46150-8_8.*
58. [^]*I. D. Dinov, "Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data," GigaScience, vol. 5, no. 1. 2016, doi: 10.1186/s13742-016-0117-6.*
59. [^]*F. S. Bashiri, A. Baghaie, R. Rostami, Z. Yu, and R. M. D'Souza, "Multi-modal medical image registration with full or partial data: A manifold learning approach," J. Imaging, vol. 5, no. 1, 2019, doi: 10.3390/jimaging5010005.*
60. [^]*M. Usama et al., "Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges," no. September, 2017, [Online]. Available: http://arxiv.org/abs/1709.06599.*
61. [^]*M. Salhov, O. Lindenbaum, Y. Aizenbud, A. Silberschatz, Y. Shkolnisky, and A. Averbuch, "Multi-view kernel consensus for data analysis," Appl. Comput. Harmon. Anal., vol. 49, no. 1, pp. 208–228, 2020, doi: 10.1016/j.acha.2019.01.001.*
62. [^]*Y. Che, C. Gneiting, T. Liu, and F. Nori, "Topological quantum phase transitions retrieved through unsupervised machine learning."*
63. [a, b, c, d, e]*H. F. Ahmad, H. Mukhtar, H. Alaqail, M. Seliaman, and A. Alhumam, "Investigating health-related features and their impact on the prediction of diabetes using machine learning," Appl. Sci., vol. 11, no. 3, pp. 1–18, 2021, doi: 10.3390/app11031173.*
64. [^]*S. Shafqat, Z. Anwar, Q. Javaid and H. F. Ahmad, "A Unified Deep Learning Diagnostic Architecture for Big Data Healthcare Analytics," 2023 IEEE 15th International Symposium on Autonomous Decentralized System (ISADS), Mexico City, Mexico, 2023, pp. 1-8, doi: 10.1109/ISADS56919.2023.10092137.*
65. [^]*S. Shafqat, Z. Anwar, Q. Javaid and H. F. Ahmad, "NER Sequence Embedding of Unified Medical Corpora to incorporate Semantic Intelligence in Big Data Healthcare Diagnostics," 09 July 2023, PREPRINT (Version 1) available at Research Square, doi: 10.21203/rs.3.rs-3148503/v1*
66. [a, b]*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.*

67. ^*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.*