Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 20s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

A Data Analytics Suite for Exploratory Predictive, and Visual Analysis of Type 2 Diabetes

Valiki Siri Vennela[1], Dr. Gachikanti swamy[2], Dr. Yeligeti Raju[3], Dr. Nalikanti Arjun[4]

[1]Computer Science and Engineering, Vignana Bharathi Institute of Technology.
Email ID: sirivennela.valiki14398@gmail.com
[2]Computer Science and Engineering, Vignana Bharathi Institute of Technology.
Email ID: Swamygachi2010@gmail.com
[3]Computer Science and Engineering, Vignana Bharathi Institute Of Technology.
Email ID: raju.yeligeti@gmail.com
[4]Computer Science and Engineering, Vignana Bharathi Institute Of Technology
Email ID: arjun.nelikanti@vbithyd.ac.in

Cite this paper as: Valiki Siri Vennela, Dr. Gachikanti swamy, Dr. Yeligeti Raju, Dr. Nalikanti Arjun, (2025) A Data Analytics Suite for Exploratory Predictive, and Visual Analysis of Type 2 Diabetes. *Journal of Neonatal Surgery*, 14 (20s), 200-212.

ABSTRACT
The availability of large volumes of electronic records of T2D patient data provides opportunities for application of big data analysis to gain insight into the disease manifestation and its impact on patients. Data science in healthcare has the potential to identify hidden knowledge from the database, re-confirm existing knowledge, and aid in personalising treatment. In this paper, we present a suite of data analytics for T2D disease management that allows clinicians and researchers to identify associations between different patient biological markers and T2D related complications. The analytics suite consists of exploratory, predictive, and visual analytics with capabilities including multi-tier classification of T2D patient profiles that associate them to specific conditions, T2Drelated complication risk prediction, and prediction of patient response to a particular line of treatment.The analyses provided in this document examine sophisticated data evaluation methods, which are possible resources for clinical and decision-making processes that may enhance the management of T2D.

Keywords: *Extensive data for medical care, data analysis, individualized treatment, medical data representation, forecasting analytics, risk assessment, type 2 diabetes*

1. INTRODUCTION

The swift progress in cloud technologies,big data frameworks,and artificial intelligence has created considerable enthusias m for creating datadriven solutions across various field
including the healthcare industry.Creating big data infrastructure, alongside data analysis for healthcare uses,necessitate me ticulous design and organization supported by strong collaboration between healthcare professionals and pertinent stakeho lders because of the delicate nature of the healthcare data involved and the potential effects it could have on patients' health

The project AEGLE, commissioned by European Union (EU), developed a big data framework aimed at providing big data services for healthcare, including electronic healthcare

record data storage, data analytics, cloud services for accelerated training of complex analytics, and real-time processing of substantial data quantities.

Figure 1 illustrates the AEGLE ecosystem. Additional information regarding the AEGLE system is available at [1]. Within the AEGLE initiative, a range of data analytics was created, including analytics for Type 2 Diabetes (T2D) among others.

T2D is a longterm illness that is becoming more prevalent worldwide. T2D is a major contributor to illness and death, resul ting in substantial utilization of healthcare resources. In 2015, Public Health England (PHE)reportedthat3.8millionpeoplein England aged over 16had diabetes[2],and it is estimated to have increased to 4.7million people in 2019 [3].T2D is a longlasting condition with rising occurrence worldwide. T2D is among the frequent causes of illness and death , resulting in considerable use of healthcare resources. In 2015, Public Health England (PHE)

alone[5].As a result, prompt intervention and successful treatment approaches are essential to lessen the impact of T2D on patient quality of life and financial expenses.

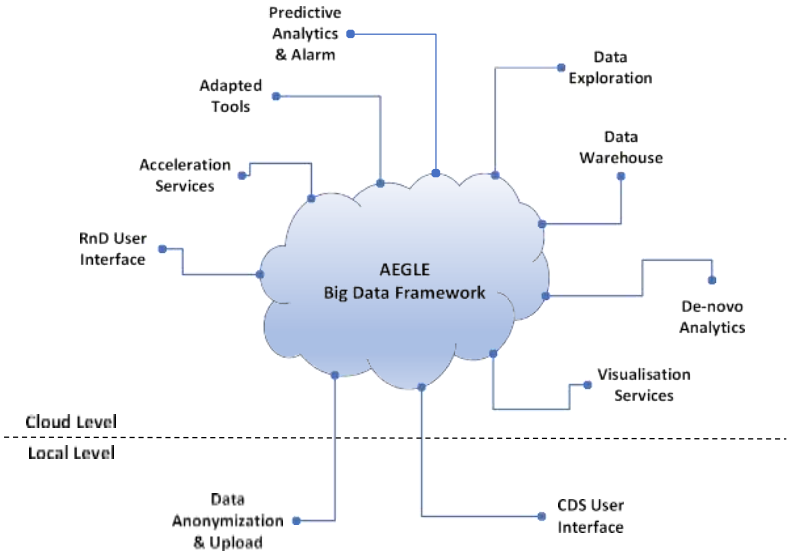


FIGURE1. Overview of AEGLE big data framework.

The rise in digital documentation of patient information over recent decades offers data analysts the chance to investigate h ealthcare databases to uncover previously unrecognized patterns and relationships that could be beneficial for enhancing th e understanding of diseases and their management.Historical data of the patient cohort enables analysts to create analytics to forecastpatient illness advancement and customize treatment plan accordingly [6].

Data analytics for T2D is a well explored topic and substantial amount of research papers can be found in the literature.A highly researched area in T2D is the prediction of complication risk. Numerous models are available, from the traditional Cox models and their variations [7]–[9] [10] to the newer machine learning approaches.models based on methods including support vector machine (SVM) [11], naives bayes[12],nearest neighbor[13],random forest[14],logistic regression [15], genetic algorithm [16] and deep learning [17]–[19].

Advancements in the evolution of data analytics for T2D create an opportunity for the creation of a tool that enables health care providers in data examination and decision-
making. In this document, we outline our research on the analysis of T2D data aimed at discovering relationships between various patient indicators and risk forecasts for different complications.and prediction of patient response to medications.The specific analyses created serve as an initial phase toward a T2D analytics suite aimed at equipping clinicia ns and researchers with insights into T2D disease and its management.

The intention of the analytics suite proposition is to emphasis the potential of using a host of data analytics as a toolbox by healthcare stakeholders for patient data analysis and decision making. An initial assessment of the feasibility of the presented analytics suite was performed as part of the

AEGLEproject's cloudenabled big data framework for healthcare data evaluation[1], [20]. Preliminary evaluations on the T 2D analytics

suite were performed by clinicians and received positive feedback.

The analysis showcased in this document is not restricted regarding innovation and clinical importance. The analytics pack age signifies preliminary efforts toward creating a structure for data analytics tools in clinical practice that will offer a new and essential diagnostic instrument.Clinic and treating T2Ddonotalwayshavesufficient information from presenting signs or symptoms to know definitively which medication or course of treatment will work. The vast variability in T2D patients and their characteristic features that can influence the course of the illness and the reaction to therapy complicate determining which kind of treatment might be most effective for each pati ent without administering specific medications to assess the response.

A tool that will cohort patients with similar risk factors for T2D will provide greatly improved indicators for what course of treatment is best, minimizing sideeffectsand enhancing treatment results with a tailored strategy. The data analysis platform outlined in this document ha s the capability to provide such a resource to healthcare professionals.

To the best of our knowledge, our paper is one of the earliest attempts towards the development of frame work for a data analytics suite for T2D.

Therest of the paper is organized as follows. In Section II an introduction to the T2D analytics suite is given. Section III describes the patient profile classifier analytic workflow. The risk prediction analytics is described in Section IV, followed by description of response prediction analytics in Section V. Section VI discusses the challenges in healthcare data analysis and Section VII concludes the paper.

ANALYTICS SUITE FOR T2D DATA ANALYSIS

This section presents the data analytics created for T2D. Figure2 demonstrates the methodology employed in developing th e proposed analytics.The methodology depicted is a common strategy utilized in data analytics development however,ongoing collaboration with clinicians at every phase, particularly during requirements gathering,analytic approach ,data collection, modeling,and feedback phases,is essential in the healthcare related data analysis process.

TheT2Danalyticssuite aims to develop exploratory, predictive, and visual analytics. The exploratory analytics focus on exploring the diabetes dataset for performing operations such as raw data pre-processing, classification, and associations between different patient markers and diabetes-related complications, and hypothesis generation. The predictive analytics focus on determining the risk faced by diabetes patients for a complication based on their biologic al markers and the likelihood of the patient experiencing a complication as time progresses in the future. The analytics addi tionally aims to foresee how diabetes patients will respond to specific treatment options and combinations. The visual analytics include custom visualizations developed for T2D data analysis that allows clinicians to gain a perceptible insight into the disease and impact on patients.

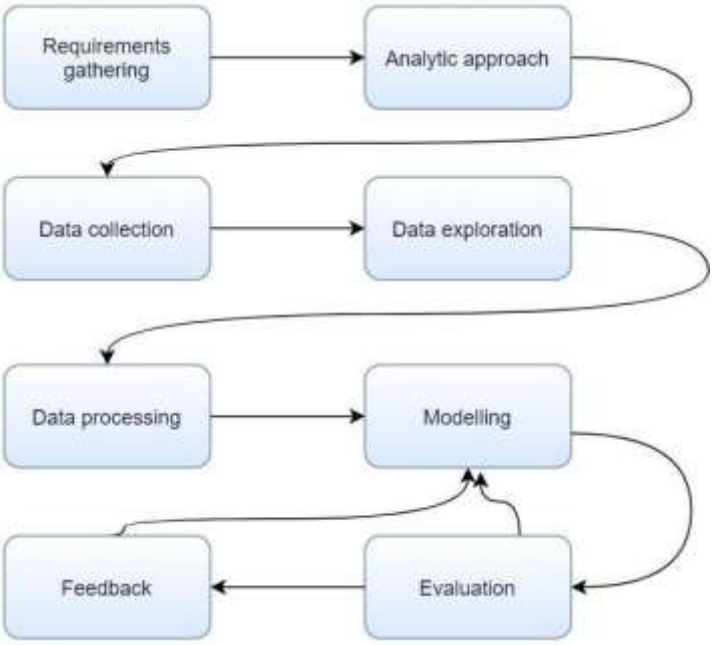


FIGURE2.Methodology followed for the development of the T2D data analytics.

T2Dpatientdataweremainlyobtainedfromtwodata sources: Croydon/Pro wellness database and Diamond database.AsummaryofthetwodatasetsisgiveninTable1. From the two databases, patient biological markers that are risk indicators of diabetes related complications capable of providing insights into patient response to treatments were identified. After extensive consultations with healthcare experts and with the aid of big data analysis techniques, several analytics were developed.

The principle approach followed for theT2Ddata analysis was to cluster the patients according to their demographics and biological markers and investigate their associations with known T2D related complications followed by the development of predictive analytics to model the associations between different patient markers. This approach helped to gain insight into hypothesis building. For instance, the example heat map in Figure 3 obtained on a synthetic dataset illustrates how associations between different variables in a dataset can be explored.

The exploratory part of the heat map analytic identifies associations between the chosen variables within the T2D database and utilises multiple statistical analysis, including correlation analysis and chi-squared analysis. The findings are then visualised on a heat map where in the associations between the variables are represented by means of endro grams.

In Figure 3 heat map, the correlation between patient bio- logical markers listed on the y-axis and three risk conditions: visual

impairments, renal replacement therapy, and death is presented. In the heat map, light shades of blue corresponds to strong correlation between the marker variables and the complication risk and darker shades correspond to weak correlations.

The clinicians and researchers can use this analytic to understand better associations between variables or con- firm already known strong and weak associations between

different variables and thus generate hypotheses for further research and analytic development.

Three T2D analytics work flows were developed, namely:patient classification system, risk assessment tool for complications, and predictor for patient treatment outcomes. Each T2D analytics procedure additionally encompasses various analyses.Details on the three analytics work flows, including its development, implementation, and findings are provided in the next three sections respectively.

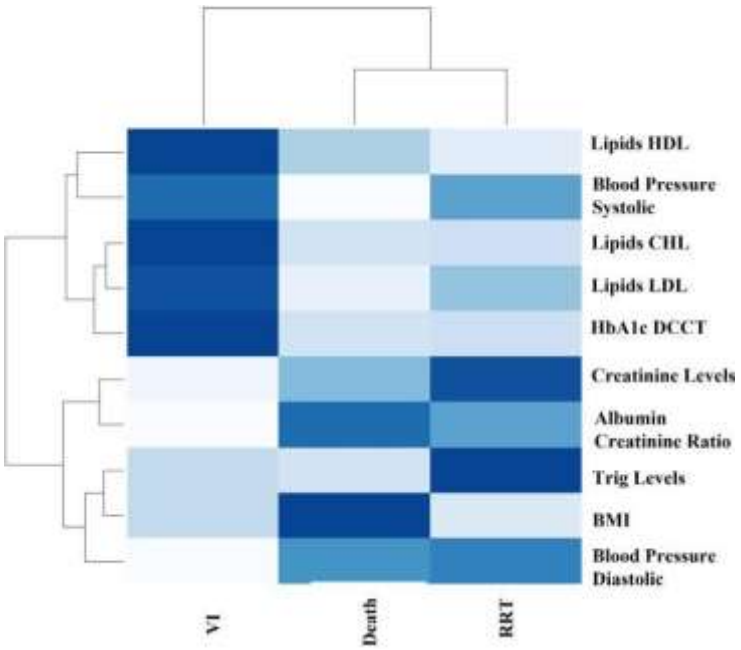


FIGURE3.Heatmapof correlation of patient marker variables(on y-axis) with three risk conditions (on x-axis).Visual Impairment,

RRT=Renal Replacement Therapy. Light colour, indicate relatively strong correlations between the variable and the risk condition.

APATIENT PROFILE CLASSIFIER

The workflow for the patient profile classifier encompasses both exploratory and visual analysis.Figure 4 depicts the eleme nts of the workflow.The aim of the classifier is to categorize patients based on a desired demographic category (e.g,age or gender) and a biological marker class(e.g.,HbA1level)and associate the classes to a complication(e.g., blindness). This multi-tier classification is achieved by means of a population pyramid analytic that helps to understand the composition of the population according to chosen criteria[21].Further, a custom visualization analytic is built based on the population pyramid analysis.

POPULATION PYRAMID ANALYTIC

In Figure 4, the exploratory segment of the analysis conducts a multistep categorization of the diabetes population starting with age groups. Subsequently, within each age group, the patient profile is additionally categorized based on patient biom arker levels such as HbA1c, Lipids, etc., and the number of patients for each age group and marker class is recorded. The t hreshold for the marker levels can be configured to a specified value by the user. Moreover, beneath each patient marker cl ass, the number of patients linked to complications such as amputation, visual impairments,etc., is collected.

TABLE1.Summary of the data sets used for developing T2D data analytics.

	Croydon/ Prowellness	Diamond
Type	Combination of database, commercial, and NHS, on secondary care of people with diabetes in South-West London/Surrey	Clean structured, commercially curated database on tertiary care of people with diabetes in Northern Ireland
Description	Parameters related to the patients and their diabetes. Longitudinal records of the patients' diabetes over several years, progression of the disease	demographics, medications, anthropomorphic markers, social and demographic factors, biochemical markers, lifestyle factors, lab results, clinical notes
Quantity	19,186 patients	16,936 patients
Mean Age (years)	60.85	61.65
Men	10432	9937
Women	8354	7155
Values at baseline:		
HbA1c recorded	12,358	10,884
Mean HbA1c (mmol/ml)	71.27	67.25
BMI recorded	17,974	15,586
Mean BMI	30.30	31.40
B.P. (systolic) recorded	15,895	12,664
Mean B.P. (systolic)	134.95	136.86

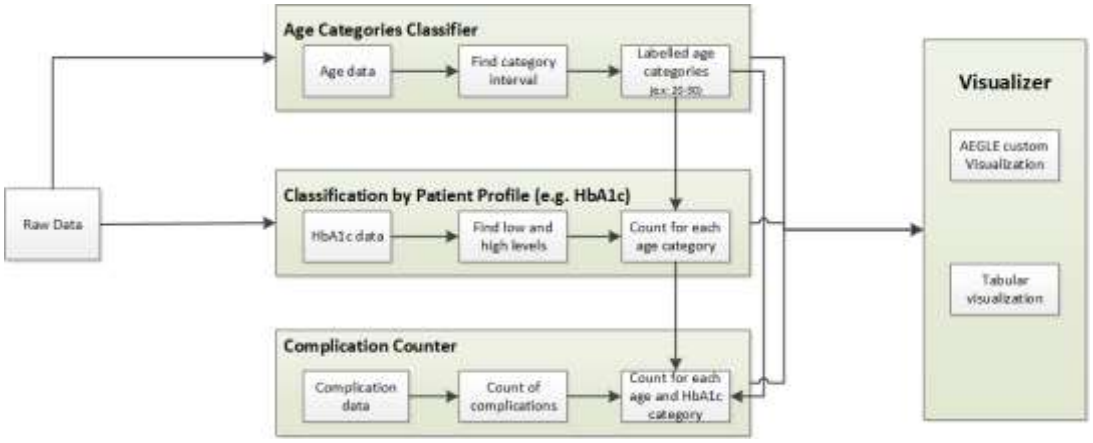


FIGURE4.Patient Profile Classifier Work flow. A two-tier classification is performed by the analytics to associate patient demographics with patient markers and T2D related complication.

The information gathered from the exploratory analysis

reveals a dual level population pyramid distribution of the

T2D patients, i.e., organized by a selected demographic category and biological marker.To visualize the two tier population pyramid, an interactive, custom visualisation analytic was developed that plots the findings of the exploratory analytic as a stacked bar-population pyramid

graph.The exploratory and visualization analysis was conducted using the R programming language, along with the rCharts [22] and High Charts [23] visualization libraries..Figure 5showsthe stacked bar population pyramid chart where the T2D patient data is categorized into age categories, low and high HbA1clevelsandthen associated with patients suffering from visual impairment complications. The visualisation helps clinicians to get an overview of the prevalence of a T2D related complication in a patient subgroup from a database.

T2D PATIENT COMPLICATION RISK PREDICTOR

T2D patients are associated with increased risk for complications such as visual impairment, coronary heart

conditions, amputations, kidney dysfunction, or cerebrovascular accident [24]. Risk assessment models are advantageous f or healthcare providers and individuals to comprehend their probability of experiencing a complication based on their exist ing T2D status.Risk assessment tools such as QRISK2 are recommended for identifying cardio vascular risk among T2D patients[25].Risk estimation for a com- plication is done by determining the factors that are likely to cause a complication risk. Identification of potential risk factors and timely intervention for control and management of the risk factors can reduce the patient's risk for T2D related complications [26]. A complication risk prediction model based on the Cox's proportional hazards model is presented in this section.

PPREDICTION ANALYTICS DEVELOPMENT

Risk assessment models for various complications such as vision loss, toe amputation, cerebrovascular incidents, cardiovas cular danger, and kidney dysfunctions were created. A cohort analysis of the patient information from the Croydon/Pro wel lness database was performed.Through consultation with clinicians and

Through literature review, numerous predictor variables with recognized risk factors for T2D-related complications were identified and encompassed the variables: age, HbA1c levels, blood pressure, BMI, and lipids. [7].

The data set consisted missing values in all of the variables. To addressing missing values, multiple imputations based on Rubin‘s rules[27]and available in the RMICE[28]package was applied. The imputation over ten iterations were applied on the database to estimate values for missing values in the database and for better complete case analysis[29].Data standardization was performed by conversion of metric units (e.g., mg/ dL to m mol/L)and merging of similar data columns. For risk prediction of a complication, the widely popular Cox‘s proportional hazards model (CPH)[30]was utilized to estimate the risk factors for each predictor variable. A censoring indicator is used to censor patients at the date of diagnosis of complications. Hence, the censoring indicator is considered to be either a complication or no complication. The survival time of the patients is computed for the patients from the time of diabetes diagnosis to the occurrence of the complication.

TABLE 2.Hazard ratios for predictor variables obtained from cph model for various complications.

Predictor variables	Hazard ratios	
	Vision impairment	Amputation
HbA1c	1.15	1.69
Blood pressure	0.99	1.52
Age	0.97	1.03
BMI	1.13	1.01
Lipids	1.20	0.53

PREDICTIONANALYTICRESULTSANDVALIDATION

Based on the pre-processed predictor variables, the CPH model computes the hazard ratios for the various predictor variables. The hazard ratios or the risk ratios indicate the extent of the risk carried by different predict or variables for a complication. Table 2 shows the hazard ratios of five predictor variables obtained for vision impairment and toe amputation risks from CPH models. The hazard ratios (HR) for a predictor variable are interpreted as follows:

- HR=1:Noeffectonthecomplication
- HR>1:Increasesriskforcomplication
- HR<1:Reducesriskforcomplication

The risk prediction models are validated using the10-fold cross-Validation method. The Croydon database was segregated to ten folds of training and test dataset. The risk models for each complication were validated separately using metrics such as sensitivity, specificity , and accuracy. The risk predictionmodelandthe10-foldcross-validationstudywas implemented in *R* programming environment. The sensitivity and specificity analysis for each fold of training and test

The dataset was examined, and the average of the metrics across the ten folds was computed. The analysis of sensitivity an d specificity occurs according to the estimation of the True Positive Rate and False Positive Rate outlined in [31] for surviv al models in risk prediction. The analysis is performed over a 15-year timeline and uses the *surv AUC R* package [32]. The *sensitivity* and *specificity* values are used to compute the prediction accuracy of the model via *Accuracy=(sensitivity)(prevalence)+(specificity)(1prevalence),*where the *prevalence* is the number of positive conditions over the total population. The outcomes of the results are presented in Table 3.

VISUAL ANALYTIC FOR SURVIVAL PROBABILITY ANALYSIS

The CPH model allows the clinicians to obtain hazard scores for the predictor variables and get a global view on their impact on the risk for a complication. It would be beneficial for clinicians and patients to obtain a risks core for individual patients

based on their biological markers. A survival analysis based visualisation analytic is designed using the CPH model developed for complication risk predictions. Two sets of survival analysis curves are presented where(1)provides a global view of a predictor variable impact on a complication risk, and(2)provides a survival probability curve for individual patients based on their conditions.

In Figure 6, different survival probability curves for each complication derived from the visual analysis are presented. Each probability curve for a complication illustrates the increase in risk for the complication overtime for all the patients considered in the Croydon/Pro wellness database. These survival curves enable clinicians and researchers to analyse complication risks for acohort group of T2D patients from a data set or geographical region and frame interventions and policies to reduce the risks of the complication. More insightful survival curves are obtained by demonstrating the impact of specific patient markers on the survival probability.InFigure7,the impact of four patient markers; HbA1clevels, body mass index (BMI), blood pressure, and high density lipoprotein(HDL)lipids levels; on the risk for visual impairment is demonstrated. Each patient marker is categorized into low, medium, and high levels along with its impact on the survival probability is shown in each individual sub-figure. For instance,inFigure7 c, the survival probability curve(in red) is higher for low BMI and the probability of survival decreases for medium BMI(green curve),and further decreases for high BMI (blue curve). Conversely, in Figure 7d for HDL lipids where high level(in mmol/L) is considered healthy, the survival probability curves progressively increases for low, medium, and high HDL lipid

levels.For an individual with T2D, comprehending their existing risk for complications can provide crucial insights for dia betes treatment and management for both clinicians and patients. Utilizing the CPH complication risk models, a survival an alysis user interface was created and implemented for possible use by healthcare providers and patients.Figure8showstheus erinterfacefora

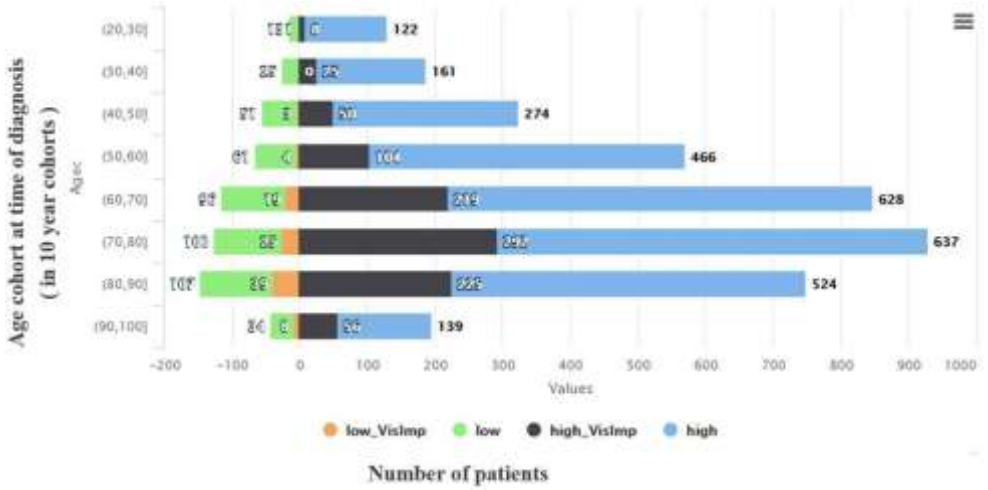


FIGURE5.Population pyramid classification. Number of patients for each category is shown. Blue corresponds to patients in the database with high

HbA1c levels (i.e. exceeding 7 mmol/L) and green relate to reduced HbA1c levels. The Black and Orange sections signify the count of patients experiencing visual impairment in the elevated and diminished HbA1c categories, respectively.

TABLE3. CPH risk prediction model validation results for complications.

Complication	Data volume	Accuracy (%)	Sensitivity (%)	Specificity (%)
Visual impairment	1220 patients	82.13	71.05	67.41
Amputation	91 patients	68.90	61.15	64.41
Cardio-vascular	144 patients	54.36	67.4	54
Renal impairment	186 patients	63.29	67	51.2

The interface enables a healthcare provider or individual to enter their biological marker levels such as BMI, HbA1C, bloo d pressure, and other parameters to generate a survival probability graph pertaining to their current state, assisting in assess ing their risk levels. This data can aid both providers and individuals in devising a plan for the management of the patient's diabetes, for example, reducing BMI or regulating blood pressure.

PATIENT TREATMENT RESPONSE PREDICTION

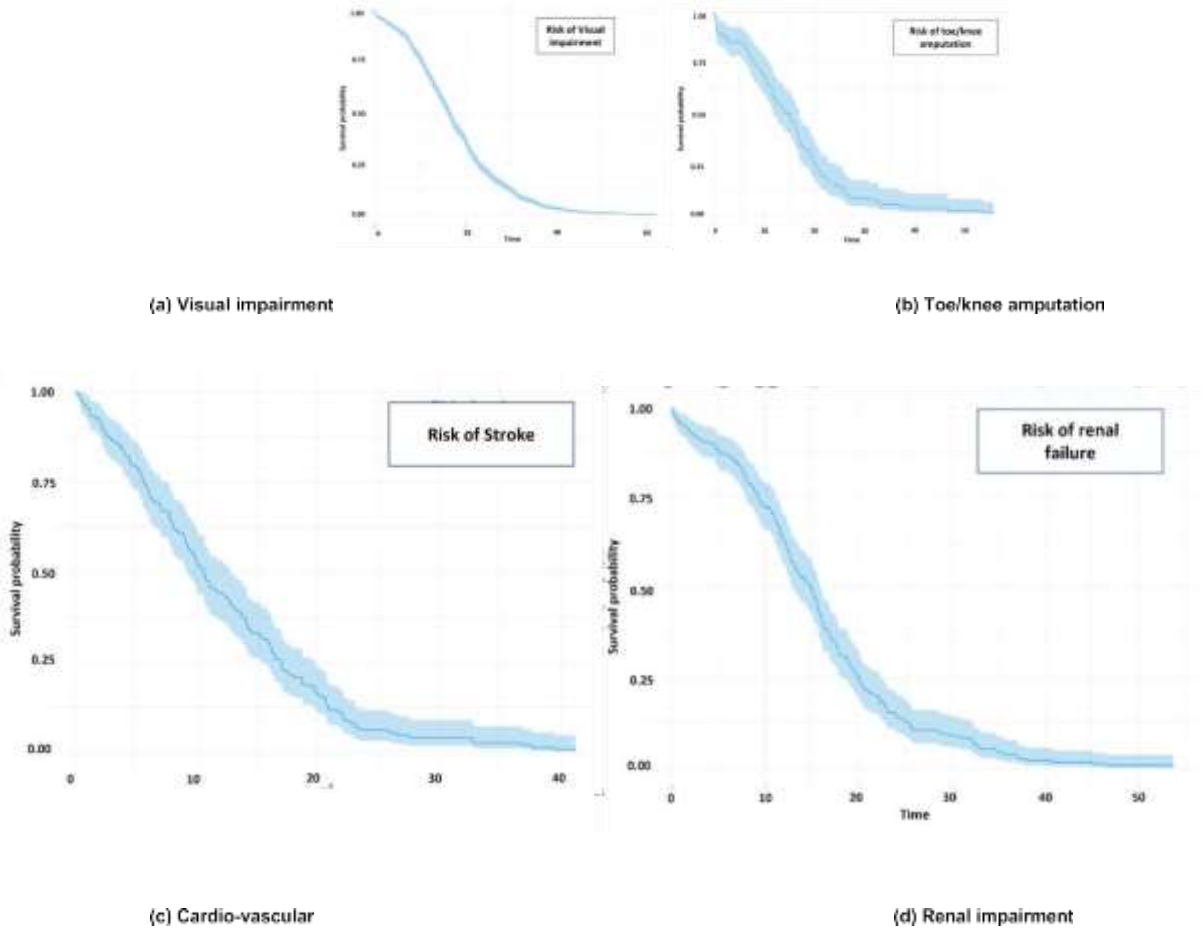
T2D is a complex condition and frequently associates with various microvascular issues such as kidney harm, vision impai rment, and neuropathy along with other major complications such as stroke and heart disease[33].Eleven distinct mechanisms are currently thought to cause diabetes,

andmore than one hundred genetic markers are linked to T2D. According to different observed traits, distinct subgroups of diabetes patients are recognized [34].

Patients with Type 2 Diabetes (T2D) from various subgroups commonly exhibit diverse responses to different categories of drugs. For example, a mix of sodium-
glucose transporter 2SGLT2 inhibitors combined with GLP1 mimetics are frequently prescribed medications for individual s with type 2 diabetes.Nevertheless, individuals with type 2 diabetes who are insulin deficient and thus susceptible to ketoa cidosis [i.e. an accumulation of keto acids in the bloodstream, a severe acute metabolic disorder] should not be administere d to patients who are prone to ketosis [20].

Identifying patientspecific factors that might cause a lack of response to certain therapies can assist in selecting the most su itable treatment category for a patient, thereby personalizing and expediting their care with established effective drugs. Ad ditional advantages include minimizing the number of medication trials, saving costs, and reducing patient exposure to pos sible adverse effects.

Big Data analysis provides opportunities in identifying groups of patients who are likely to respond to a specific line of treatment. Patient cohort data over an extended period and a combination of dataset sources enables to detect patterns and patient response to medications over time. Analysis of cohort data provides the potential to build prediction models to predict patient response to specific medications based on patient characteristics. The Diamond database offers a cohort



(a) Visual impairment (b) Toe/knee amputation

(c) Cardio-vascular (d) Renal impairment

FIGURE6.Survival probability curves show the rate of increase in risk for complication with time. Y-axis indicates the survival probability and X-axis is duration in years.

of T2D patients, encompassing medications advised for them. A prediction model based on machine learning is developed to forecast patient reactions to third-

line agents (i.e., SGLT2 inhibitors). The SGLT2 inhibitors are utilized to encourage glycosuria and are a sanctioned catego ry of medications in the management of T2D.

TREATMENT RESPONSE PREDICTOR IMPLEMENTATION AND RESULTS

Prior to developing the prediction system, it is essential to train the machine learning algorithm to understand how to forec ast patient reactions to SGLT therapy.From the T2D Diamond database, patients treated with SGLT are considered to show good response when they show improvement in HBA1 clevels by11mmol/L over three or six month period. Based on the

HbA1c changes, the patient response is classified into the

upper and lower extremes of favorable response and unfavorable response classifications. Subsequently, characteristics of t he patient are identified that may likely affect their reaction to treatments.The features selected include age, gender, duration of diabetes, weight, BMI, HbA1 clevels, and medications taken. A cohort data of approximately 2300 patients from the Diamond database was selected to train and build the prediction model.Outofthe2300 patients, approximately1000 patients belonged to the good response category and the remaining to the bad response category. The dataset was further classified into a training dataset (80%)and a validation dataset(20%).

The renowned machine learning algorithm support vector machine (SVM) was selected as our forecasting model because o f its proven high-performance accuracy in data prediction[35].

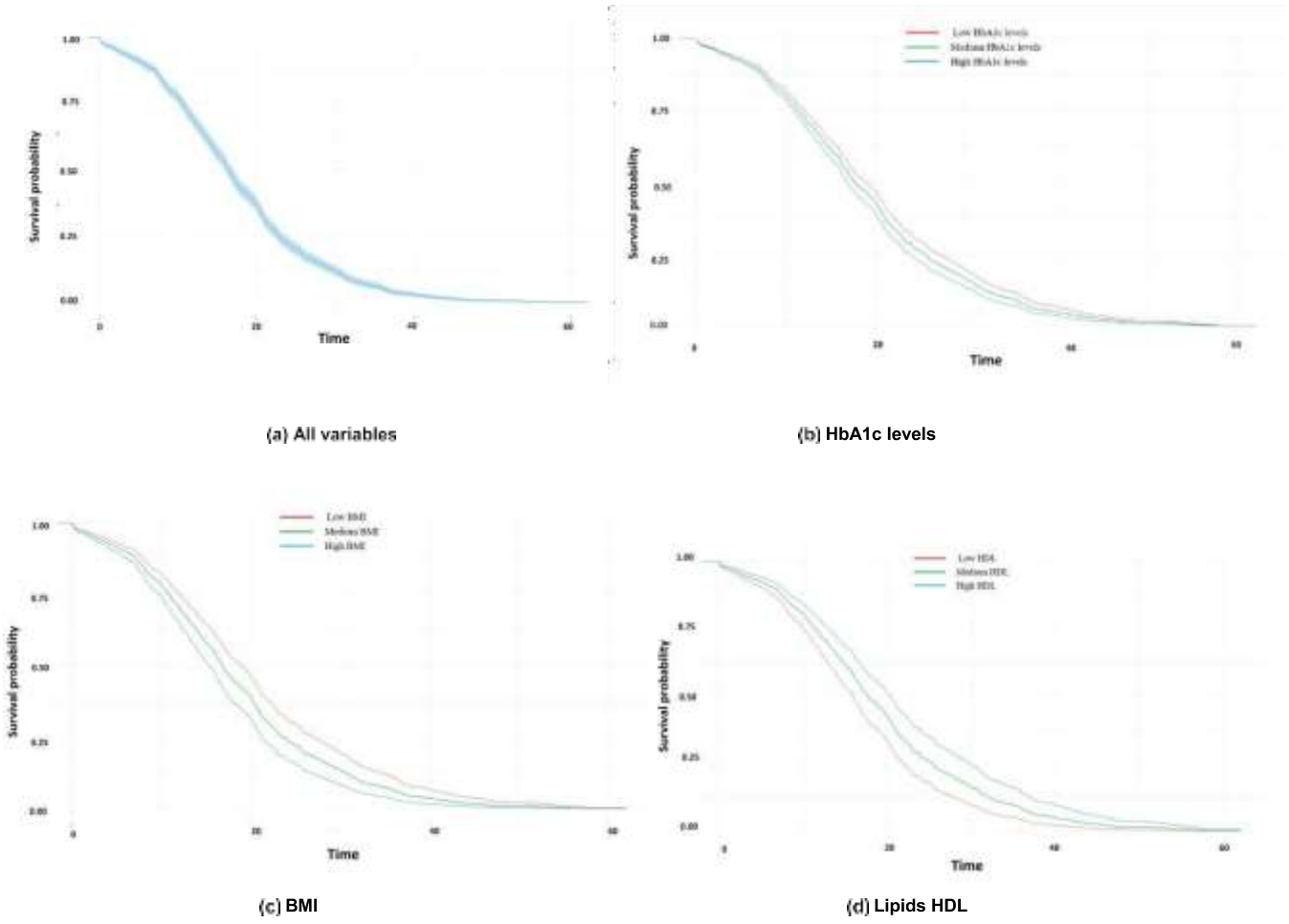


FIGURE7:Survival probability graphs for visual impairment depicting the influence of predictor variables on the risk rate. The yaxis represents the Survival probability, while the xaxis denotes duration in years. (a) illustrates the survival probability graph when the risk of all patient biomarkers is taken into account. (b) depicts the survival graph for low, medium, and high levels of H bA1c. (c) presents the survival graph for low, medium, and high BMI levels. (d) shows survival graphs for low, medium, and hig h HDL levels. It can be noted that the survival graph produced by the analysis is diverse and reflects the effect of patient biomar kers on the risk of visual impairment.

The prediction model was implemented in the R programming environment. The SVM model was validated using a 5-fold Cross validation analysis, reaching an average forecasting accuracy of 65.05%.was obtained in the 5 fold cross validation, and the best prediction accuracy obtainedwas73.3%.The out comes of the prediction model are shown in Table 4.

TABLE-4.Overview of the SVM prediction model regarding patient reactions to the SGLT treatment line.

	Outcome
Patients (n)	2300
Validation method	5-fold cross-validation
Best accuracy	73.30%
Sensitivity	66.85%
Specificity	75.21%
Average accuracy	65.05%

The treatment response prediction model provides capabilities to analyse the response of patient subgroups to a specific line of treatments. This leads to a reduction in medication trials to find the most effective treatment for T2D patients.

Integrating these analytics proves advantageous for healthcare practitioners to anticipate how patients with specific charact eristics will react to a given treatment approach. As the accumulation of data on patient reactions to therapies continues to grow, the presented prediction model can be further periodically trained to be more reliable and to give accurate predictions. In addition to models such as SVM, the presented approach can be adapted to include other machine learning models such as Naives Bayes and k-nearest neighbor (kNN) to enhance assistance for clinicians in making treatment choices.

2. DISCUSSION- CHALLENGES IN MEDICAL DATA ANALYSIS

Creating data analytics for the healthcare sector involves distinct difficulties stemming from the delicate nature of the infor mation and the significant implications of the results. Below are some prevalent obstacles in data analysis accompanied by suggestions to address these issues.

- Data quality: Like most big data solutions, data quality is a problem too in the health care domain. Main issues with data quality arise due to lack of completeness in the data (missing values), data repetition, irregular and

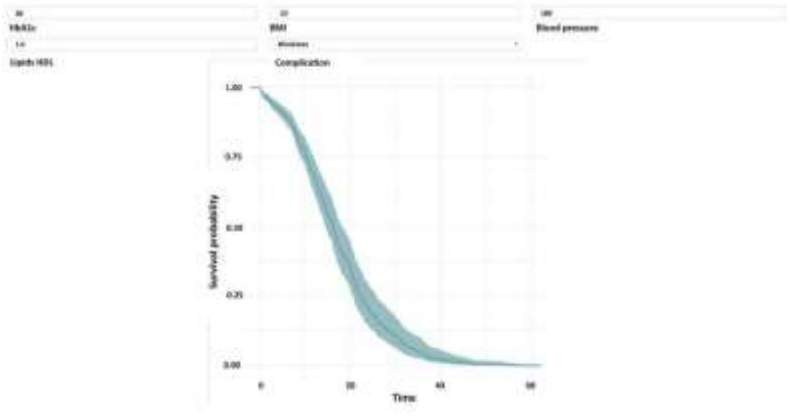


FIGURE8.Interactive survival likelihood graph for evaluating patient complication risk. A healthcare provider inputs patient in dicators such as HbA1c., BMI and selects a complication and obtains a patient specific survival probability curve for the chosen complication.

inconsistent data refresh, lower precision and erroneous data entries. Data cleansing techniques such as numerous imputati

ons and normalization of various data sources can tackle data integrity problems.In our research, various methods to enhan ce data quality were implemented. For instance, multiple imputations were part of data preprocessing for predictors of complication risk detailed in Section IV. The Diamond database and Pro wellness databases were frequent ly combined to ensure adequate data volume for our analytics. Data standardization was handled through the transformation of metric units (e.g., mg/dL to m mol/L) and the consolidation of analogous data colum ns. Feasibility:

Creating big data analytics for healthcare necessitates a collaborative approach between data analysts and clinical specialist s. Before developing data analytic solutions, a feasibility assessment is essential. From the clinician's viewpoint, use case s cenarios must be formulated to illustrate the pertinent clinical issue at hand. Clinicians must also evaluate the data feasibilit y, meaning the accessibility and quality of the data. For example, during our research, the creation of sophisticated deep lea rning
based data analytic solutions, while desired, was not viable due to the unavailability of large datasets. From the perspective of data analysts, technical feasibility analyses will enable stakeholders to reach a viable solution for possible development. Aspects such as cost-effectiveness, clinical relevance, and application of out comes in clinical trials must be evaluated before investing in development of data analytics.

Solutions for Decision Support System: It is essential for

the strategies created to serve as a Decision Support System (DSS). A DSS would act as an enabling resource for clinicians to make decisions related to patient care .For instance, the survival probability analyser UI presented in Section IV can be utilized by

to assess a specific patient's current threats to an issue and utilize the information to plan on disease management. Further modifications to the probability analyser to enable it to describe a patient risk in evaluative form (e.g., low, moderate, and elevated risk) might be utilized by a patient to determine their risk classificatio n either autonomously or with assistance from a healthcare provider.

Simplicity:
Often, many data analytic challenges do not necessitate Big Data infrastructure and can be addressed with existing data mi ning tools. However, with the escalation of both the volume and accuracy of healthcare records, advanced big data tools be come essential. Despite the complexity and sophistication involved in developing big data analytics, it is preferred that the outcomes of these analytics yield simple solutions that can be employed in routine clinical practice. One of the goals of the diverse analytics presented in this document is to provide straightforward solutions for comprehending and enhancing T2D disease management. We believe the data analytics outlined in the paper, with additional clinical validation, hold promise for integration into clinical practice, particularly for tasks such as data visualization and risk assessment. A critical factor b eing that it

Does not require significant IT infrastructure and can be adapted according to the data availability and requirements of the clinician.

Extensibility: A desirable feature for data analytics is to permit inclusion of additional features and models in analysis. The analytics presented in the paper can be extended to include relevant features for analysis. For instance, the population pyramid analyser and complication risk predictor can be extended to include new biological markers for analysis and are not restricted to the markers presented in the paper. As discussed in Section V-A, the presented prediction model approach is not limited to SVM model alone and can be extended for use with other machine learning models.

Scalability: A key characteristic for analytics is the ability to scale the analysis to larger data sets. The presented analytics are tested on relatively smaller datasets with lower than 20,000 patients. However, further tests are required to evaluate the capability for analyzing significantly larger datasets.

The tool requires further testing of the analytics on a large scale using more external T2D databases. Planned future works include design of a robust framework for the analytics suite that includes flexibility for clinicians to choose from multiple models. Further, the data models in the analytics will be extended to include more advanced, clinically validated models.

3. CONCLUSION

In this paper, we presented an analytics suite that performs exploratory, predictive, and visual analysis of T2D data. Three types of analytics workflows were presented that perform: (1) classification of T2D patients into required categories and identifying associations to a condition of interest,

(2) analysis of T2D database to build a predictive model that can assess risk of patients toT2D related complications, and (3) prediction of patients' response to a specific line of treatment plan. The visual analytics provides a simplified representation of the outcome for clinicians and patients.

The analytics presented have the potential to support clinicians to decide treatment plans for T2D patients. This offers huge advantage that had not been previously possible for a more personalized approach to treating T2D that will be safer and more beneficial for the patient as it will minimise side effects and offer faster, more effective treatment. It will also provide economic advantages to the healthcare system.

Possibilities for future work include building and training the model on larger databases to increase the prediction accuracy and develop more robust prediction models by adopting artificial intelligence methods, and clinical validation of the data analytics.

REFERENCES

[1] D.Soudris,S.Xydis,C.Baloukas,A.Hadzidimitriou,I.Chouvarda,

K. Stamatopoulos, N. Maglaveras, J. Chang, A. Raptopoulos, D. Manset,and B. Pierscionek, ‘‘AEGLE: A big bio-data analytics framework forintegrated health-care services,’’ in Proc. Int. Conf. Embedded Comput.Syst., Archit., Modeling, Simulation (SAMOS), Jul. 2015, pp. 246–253.

[2] N. Holman, B. Young, and R.Gadsby, ‘‘Currentprevalence oftype 1 andtype2diabetesinadultsandchildrenintheU.K.,’’DiabeticMed.,vol.32,no. 9, pp. 1119–1120, Sep. 2015.

[3] NumberofPeopleWithDiabetesReaches4.7Million.Accessed:Oct. 30, 2019. [Online]. Available: https://www.diabetes.org.U.K./about_us/news/new-stats-People-livingwith-diabetes

[4] C.D.MathersandD.Loncar,‘‘Projectionsofglobalmortalityandburdenofdiseasefrom2002to2030,’’PLoSMed.,vol .3,no.11,p.e442,Nov.2006.

[5] AmericanDiabetesAssociation,‘‘Economiccostsofdiabetesinthe U.S. in 2017,’’ Diabetes Care, vol. 41, no. 5, pp. 917–928, 2018, doi:10.2337/dci18-0007.

[6] J.M.M.Rumbold,M.O’Kane,N.Philip,andB.K.Pierscionek,‘‘Bigdataanddiabetes:Theapplicationsofbigdatafordi abetescarenowandinthefuture,’’ Diabetic Med., vol. 37, no. 2, pp. 187–193, Feb. 2020.

[7] J. Hippisley-Cox and C. Coupland, ‘‘Development and validation of riskprediction equations to estimate future risk of blindness and lower limbamputationinpatientswithdiabetes:Cohortstudy,’’ BMJ,vol.351,no.1,Nov. 2015, Art. no. h5441.CI.Marzona,F.Avanzini,G.Lucisano,M.Tettamanti,M.Baviera,

[8] A. Nicolucci, and M. C. Roncaglioni, ‘‘Are all people with diabetes andcardiovascular risk factors or microvascular complications at very highrisk? Findings from the risk and prevention study,’’ Acta Diabetolog.,vol. 54, no. 2, pp. 123–131, Feb. 2017.

[9] S.Basu,J.B.Sussman,S.A.Berkowitz,R.A.Hayward,andJ.S.Yudkin,‘‘Developmentandvalidationofriskequations forcomplicationsoftype2diabetes(RECODe)usingindividualparticipantdatafromrandomisedtri- als,’’LancetDiabetesEndocrinol.,vol.5,no.10,pp.788–798,Oct.2017.

[10] E. B. Schroeder, S. Xu, G. K. Goodrich, G. A. Nichols, P. J. O’Connor,and J. F. Steiner, ‘‘Predicting the 6- month risk of severe hypoglycemiaamongadultswithdiabetes:Developmentandexternalvalidationofapre- dictionmodel,’’J.DiabetesComplications,vol.31,no.7,pp.1158–1163,Jul.2017.

[11] N. Barakat, A. P. Bradley, and M. N. H. Barakat, ‘‘Intelligible supportvector machines for diagnosis of diabetes mellitus,’’ IEEE Trans. Inf.Technol. Biomed., vol. 14, no. 4, pp. 1114–1120, Jul. 2010.

[12] B. Liu, Y. Li, S. Ghosh, Z. Sun, K. Ng, and J. Hu, ‘‘Complication riskprofiling in diabetes care: A Bayesian multi-task and feature relation-ship learning approach,’’ IEEE Trans. Knowl. Data Eng., vol. 32, no. 7, pp.1276–1289,Jul.2020.

[13] A.PavateandN.Ansari,‘‘Riskpredictionofdiseasecomplicationsintype2diabetes patientsusingsoft computingtechniques,’’ in Proc.5th Int. Conf.Adv. Comput. Commun. (ICACC), Sep. 2015, pp. 371–375.

[14] J.Yan,X.Du,Y.Yu,andH.Xu,‘‘Establishmentofriskpredictionmodelfor retinopathyin type 2diabeticpatients,’’ in Proc. Int.Conf. Brain Inform.Haikou, China: Springer, 2019, pp. 233–243.

[15] A.Dagliati,S.Marini,L.Sacchi,G.Cogni,M.Teliti,V.Tibollo,P.D.Cata, L. Chiovato, and R. Bellazzi, ‘‘Machine learning methods to predict dia-betes complications,’’ J. DiabetesSci. Technol., vol. 12, no.2,pp.295–302,Mar.2018.

[16] K. V. Dalakleidi, K. Zarkogianni, V. G. Karamanos, A. C. Thanopoulou,and K. S. Nikita, ‘‘A hybrid genetic algorithm for the selection of thecritical features for risk prediction of cardiovascularcomplicationsintype 2 diabetes patients,’’ in Proc. 13th IEEE Int. Conf. Bioinf. BioEng.,Nov. 2013, pp. 1–4.

[17] R. GargeyaandT.Leng,‘‘Automatedidentificationofdiabeticretinopathyusingdeeplearning,’’Ophthalmology, vol. 124,no.7,pp.962–969,2017.

[18] M.-H.Hsieh,L.-M.Sun,C.-L. Lin,M.-J.Hsieh,K.Sun,C.-Y.Hsu,sA.-K. Chou, and C.-H. Kao, ‘‘Development of a prediction model forcolorectalcanceramongpatientswithtype2diabetesmellitususingadeepneural network,’’ J. Clin. Med., vol. 7, no. 9, p. 277, Sep. 2018.

[19] Y.Cheng,F.Wang,P. Zhang,andJ.Hu,‘‘Riskpredictionwithelectronichealthrecords: Adeeplearningapproach,’’in Proc.SIAMInt.Conf.DataMining, Jun. 2016, pp. 432–440.

[20] D.Masouros,K.Koliogeorgi,G.Zervakis,A.Kosvyra,A.Chytas, S. Xydis, I. Chouvarda, and D. Soudris, ‘‘Co-design implications of cost-effective on-demand acceleration for cloud healthcare analytics: TheAEGLE approach,’’ in Proc. Design, Autom. Test Eur. Conf. Exhib.(DATE), Mar. 2019, pp. 622–625.

[21] M. Richmond, Population Pyramids. Corvallis, OR, USA: Oregon StateUniv.,2014.

[22] rChartsbyRamnathVaidyanathan.Accessed:Oct.25,2019.[Online]. Available:https://ramnathv.github.io/rCharts

[23] Highcharts. Accessed: Oct. 25, 2019. [Online]. Available: https://www.highcharts.com/demo/bar-negative-stack.

...