Improving Accuracy for Diabetes Mellitus Prediction by Using Deepnet

Riyad Alshammari¹, Noorah Atiyah², Tahani Daghistani¹, Abdulwahhab Alshammari¹

¹Health Informatics Department, College of Public Health and Health Informatics

King Saud Bin Abdulaziz University for Health Sciences (KSAU-HS)

King Abdullah International Medical Research Center (KAIMRC)

Ministry of National Guard Health Affairs, Riyadh, KSA

² Faculty of Health Sciences, Simon Fraser University, Burnaby British Columbia, Canada

Abstract:

Diabetes is a salient issue and a significant health care concern for many nations. The forecast for the prevalence of diabetes is on the rise. Hence, building a prediction machine learning model to assist in the identification of diabetic patients is of great interest. This study aims to create a machine learning model that is capable of predicting diabetes with high performance. The following study used the BigML platform to train four machine learning algorithms, namely, Deepnet, Models (decision tree), Ensemble and Logistic Regression, on data sets collected from the Ministry of National Guard Hospital Affairs (MNGHA) in Saudi Arabia between the years of 2013 and 2015. The comparative evaluation criteria for the four algorithms examined included; Accuracy, Precision, Recall, F-measure and PhiCoefficient. Results show that the Deepnet algorithm achieved higher performance compared to other machine learning algorithms based on various evaluation matrices.

Keywords: Diabetes, Artificial Intelligence, Deep Learning

*Corresponding Author: riyadalshamamri@gmail.com

DOI: 10.5210/ojphi.v12i1.10611

Copyright ©2020 the author(s)

This is an Open Access article. Authors own copyright of their articles appearing in the Online Journal of Public Health Informatics. Readers may copy articles without permission of the copyright owner(s), as long as the author and OJPHI are acknowledged in the copy and the copy is used for educational, not-for-profit purposes.

Introduction:

Diabetes is a severe disease that affects all genders and ages [1]. Diabetes is a metabolic and systemic disease in which a disruption in the metabolism of carbohydrates occurs because of insufficient insulin production for the body's metabolic needs [2]. There are two main types of diabetes; Type 1, or insulin-dependent diabetes, which is a result of the elimination of insulin-producing pancreatic cells [2]. Type 2, or non-insulin-dependent diabetes, correlates to obesity and
results from a relative lack of insulin [2]. This disease is a result of an individual's carbohydrate intake exceeding the capacity of their pancreas's production of insulin [2]. The gravity of the condition is evident in the form of complications [3]. Common complications of diabetes include, but are not limited to, heart disease, stroke, and kidney disease, which can result in higher mortality [3]. At the patient level, an individual may fail to recognize they have the disease or fail to receive prompt appropriate care resulting in poor prognosis [3].

The global prevalence of diabetes for adults aged more than 18 years old was 8.5% in 2014 per the World Health Organization (WHO) report [4]. Of the population impacted by diabetes, 80% of the people lived in low-income and middle-income countries, with the highest diagnosis being Type 2 diabetes; however, there is an alarming rise in the prevalence of both Type 1 and 2 diabetes [1]. Parallel to the increasing prevalence rate of the disease, there is an increase in associated consequences due to the complications of diabetes, i.e. increase in heart disease, stroke, and poor health [3]. Therefore, mortality rates as a result of diabetes and its comorbid health problems are rising proportionally [5]. In 2015, there was an estimate of 1.6 million deaths as a direct cause of diabetes [1]. The International Diabetes Federation reported that the disease affects one in 11 adults worldwide, with one person dying of the disease every six seconds [1]. In 2030, WHO anticipates that diabetes will be the seventh leading cause of death [4]. In Saudi Arabia, there is an excessive prevalence of diabetes, with an estimated rise of more than 2.5 million patients by 2030 having the disease [6].

Early prediction of Type 2 diabetes is a prominent health research topic in Saudi Arabia. Diabetes Risk Score was the most convenient tool for diabetes prediction [7]. However, this method needs human intervention in decision-making. Nowadays, Computational models to predict the risk of diabetes can significantly support healthcare providers with decision-making and assist self-disease management, which, in turn, can potentially decrease the diseases associated mortality rates [8]. Therefore, machine learning is gaining attention in the health field as these techniques produce high performance in predicting diabetes.

Specifically, these models can help identify those who are at high risk of having diabetes, and for which early prevention and control programs can improve health outcomes [7,9]. At the same time, these techniques reduce the human error in necessary healthcare decisions. Thus, decreasing health burden and utilizing health service resources [5]. Ideally, further development of models that incorporate prior knowledge would be promising for diabetes prediction [10]. The availability of a patient's health data could help to extract meaningful information and hidden knowledge to better the prognosis of the individuals affected by this disease.

**Background**

**The Biology of Diabetes**

Type 1 diabetes is an abnormal immune reaction controlled by a portion of the HLA-D region genes and works directly against molecules expressed only on the β-cells [11]. The pathway for immunological response systems is complex but involves mounting a response towards foreign antigens [11]. In Type 1 diabetes, similar attacks occur on certain pancreatic β-cells resulting in an insulin deficiency [11].
Type 2 diabetes is one of the most frequent metabolic disorders, and it is a heterogeneous disease distinguishable by its deficient production of insulin secretion via the pancreatic islet β-cells [12]. Insulin deficiency results in insulin resistance or impaired insulin sensitivity, which leads to a decline in patient health [12].

Genome-wide association studies found for islet function; more than 400 Type 2 diabetes-associated gene variations influenced secretion [12] [13]. However, genetic roles in the individual genes explain less than 20% of overall diabetic disease risk [12] [13]. In contrast, the literature on lifestyle modification indicates sedentary lifestyles, poor diet, and a myriad of social determinants of health (such as; low socio-economic status, psychological conditions, poor environment) all play a predominant role in the development of Type 2 diabetes [12] [14] [15] [16]. Furthermore, parental lifestyle has longitudinal impacts on the life course of an individual. Within utero programming and early postnatal metabolic transformation correlates to the risk of diabetes due to DNA methylation [12] [17] [18].

Type 2 diabetes results from a variety of factors but is often mitigated through lifestyle changes and preventative measures such as diet change, increased exercise, and overall holistic integration of health.

In summary, over the past 50 years, diabetes mellitus, or diabetes in layman's terms, continues to increase, with individuals in Western, Western Pacific, Asian and African countries all experiencing an increase in disease prevalence [12]. Cho and colleagues [19] predict globally for years 2017 to 2045, a diabetes rate increase of at least 50%, meaning approximately 693 million people will be affected by the disease creating an estimated healthcare cost of US$850 billion per year.

**Diabetes in Saudi Arabia**

In the past four decades, Saudi Arabia has undergone significant socio-economic change [20]. Specifically, Saudi Arabia has seen an increase in an ageing population, progressive urbanization, decreased infant mortality rates and increased life expectancy [20]. The changes in population demographics also couples with a rapid change in lifestyles, where individuals are moving towards westernized patterns of consumption, shown in changes in nutrition, less physical activity, higher rates of obesity, and increases in smoking—all resulting in a dramatic rise in the prevalence of diabetes [21] [22] [23] [24].

The WHO reported in 2016 [25], the prevalence of diabetes in Saudi Arabia was 14.7% for males and 13% of females. The WHO [25] also found high prevalence of overweight individuals (67.5% males; 69.2% females), obesity (29.5% males; 39.5% females), and inactivity (52.1% males; 67.7% females). The WHO [25] further reported high mortality rates attributed to diabetes with 1070 males and 500 females (aged 30-69) and 1460 males and 1020 females (aged 70+) dying due to the disease. Overall, diabetes is an important health concern for the citizens of Saudi Arabia. Integrating early detection and prediction models would have both national and global benefits.
AI and Diabetes

Dalakleidi et al. [26] applied Evolving Artificial Neural Networks (EANNs), Bayesian-based algorithm, decision trees and Logistic Regression to predict the progress of diabetes and its complications related to cardiovascular disease. They achieved an accuracy of 80% with the EANNs algorithm. For the difficulties, they produced an accuracy of 92.86%. Meng et al. [27] compared three different techniques, namely Logistic Regression, decision tree and Artificial Neural Networks, to predict diabetes and prediabetes. They achieved the best performance with a decision tree with an accuracy of 77.87%, a sensitivity of 80.68% and specificity of 75.13%. Wang et al. [28] built a classification model to recognize people of developing Type 2 diabetes. They compared Artificial Neural Networks (ANNs) and Multivariate Logistic Regression (MLR). They showed that ANN outperformed MLP. Research is promising and demonstrates that AI has the potential to help in the diagnostic framework of diabetic or prediabetic patients.

Methods

The following section discusses the methodology of this research article. Furthermore, this section describes the gathering of the data-set and feature information; while also explaining the algorithms used in the following research and evaluation criteria.

A. Data-set and Features

The collection of health data-sets were between the years 2013 and 2015. The health-data was from the Electronic Health Record of the Ministry of National Guard Health Affairs databases for all adult patients who had tested for Hemoglobin A1c (HgbA1c). The process of labelling patients as diabetic relied on the results of the HgbA1c. If the value of HgbA1c was higher or equal to seven, patients were classified as diabetic. If the value of HgbA1c was less than seven, patients were classified as non-diabetic. After the pre-processing of the data-sets, the exclusion criteria (exempting participants from further analysis) included those with a missing value of 40% and higher. The usage of the manual inspection and domain knowledge technique allowed researchers to remove implausible values. Furthermore, to check the quality of data, this study used R to analyze the given information. Table 1 shows a descriptive analysis of the attributes. The data sets have 17 attributes organized into three categories:

1) Demographic attributes such as gender, age, and region;
2) Measurement attributes such as the Body Mass Index (BMI) and blood pressure;
3) Lab tests.

<table>
<thead>
<tr>
<th>Risk Factors</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cities</td>
<td></td>
</tr>
<tr>
<td>Riyadh</td>
<td>54141 (81.63%)</td>
</tr>
<tr>
<td>Dammam</td>
<td>11085 (16.71%)</td>
</tr>
<tr>
<td>Jeddah</td>
<td>1099 (1.66%)</td>
</tr>
</tbody>
</table>
Males represented 55.50% of the data-set, while females represented 44.50%. Most of the data belonged to patients aged 45 to 84 years old. The percentage of diabetic patients in the data-set was 64.47%. The incidence of diabetes for both genders was higher in those aged 65-84 years, with males at 47.83%, and females at 48.6%. Comparatively, those with an age range of 45-64 years demonstrated the following: 37.89% male and 38.03% female. The results show that the
female patients in the data-sets had a higher Body Mass Index (BMI) and blood pressure measurement compared to the male patients, see Table1.

B. Algorithms and Evaluation Criteria

BigML [29] is a cloud computing service that provides Machine Learning as a Service (MLaaS). The BigML offers a collection of state-of-the-art machine learning algorithms and demonstrates the ability to solve real-world challenges in various domains. The BigML was used to build four Machine Learning-based algorithms, namely Ensemble, Models (decision trees), Deepnet, and Logistic Regression.

A model in BigML is similar to decision tree representation. Each node represents one of the input attributes (predictors), with the first node being the root. Each node except the root has two branches (leaves) that represent a value of an attribute. A leaf represents the outcome of the class (objective field) in the chain of branches, starting from the root to the leaf end.

An ensemble in the BigML is a group of machine learning algorithms joined together to make a more reliable model. Logistic Regression is a supervised machine learning algorithm that uses a logistic function with the input values to build a learning model. Deepnet in the BigML is an optimized form of deep neural networks that is suitable for classification problems. The Deepnet is a supervised machine learning model that simulates the human brain neural circuitry.

A 10-fold cross-validation technique was applied to evaluate the performance of each machine learning algorithms. It works by dividing the data set into ten equal folds. The training of the machine learning model utilized a one-fold test on the remaining folds, with an iteration of ten. At the end of the tenth iteration, the result shows the average of all the ten folds [30].

The application of the following matrices selected the best model in predicting the label classes (diabetic vs. non-diabetic); True Positive rate, False Positive rate, Precision, Recall, Area Under the Curve and F-measure. The calculated metrics were:

- **Accuracy**: which represents the number of correctly classified records over the total number of evaluated records, calculated based on equation 1:
  \[ \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \]  
- **Precision**: which represents the number of true positives correctly identified as diabetic patients over the total number of positive predictions, calculated based on equation 2:
  \[ \text{Precision} = \frac{TP}{TP + FP} \]  
- **Recall**: which represents the number of true positives correctly identified as diabetic patients over the total number of positive records, calculated based on equation 3:
  \[ \text{Recall} = \frac{TP}{TP + FN} \]  
- **F-measure**: which represents the harmonic mean of precision and recall, calculated based on equation 4:
  \[ 2 \times \left( \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \right) \]  
- **PhiCoefficient**: represents the Matthews Correlation Coefficient, calculated based on equation 5:
\[
\Phi(\text{Coefficient}) = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}
\]  

(5)

**Results and Discussion:**

The aim of this study was the following: to build a machine learning model(s) that can predict diabetes with high accuracy. Therefore, the usage of the BigML machine learning platform helped in the creation of the four machine learning models, namely Ensemble, Models (decision tree), Deepnet and Logistic Regression. Each machine learning algorithm has different machine learning techniques. The overall goal of this study was to find the best performance model and apply its technique to predict diabetes. The performance of the Deepnet model was better than Models (decision tree), Ensemble and Logistic Regression on all of the evaluation criteria, Table 2.

**Table 2: Evaluation of Predicting diabetes using AI Techniques.**

<table>
<thead>
<tr>
<th></th>
<th>ENSEMBLE</th>
<th>MODELS</th>
<th>Deepnet</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>88.1</td>
<td>87.8</td>
<td>88.48</td>
<td>88.19</td>
</tr>
<tr>
<td>Precision</td>
<td>87.9</td>
<td>87.7</td>
<td>88.29</td>
<td>88.38</td>
</tr>
<tr>
<td>Recall</td>
<td>87.8</td>
<td>87.6</td>
<td>88.36</td>
<td>87.63</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.8783</td>
<td>0.8761</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>PhiCoefficient</td>
<td>0.7566</td>
<td>0.7522</td>
<td>0.77</td>
<td>0.76</td>
</tr>
</tbody>
</table>

The prediction of diabetic patients who may not know they have the disease is a crucial challenge in the healthcare domain. The machine learning technique demonstrates the ability to predict diabetes with high accuracy using only 17 attributes. Furthermore, an offered perk of this method is information collection can occur from routine checkups at a healthcare clinic. This process will allow the integration of up-to-date information into the system expediting medical care and easing the burden on healthcare workers and patients.

Changing the healthcare workflow can enhance the early healthcare assessments of those with diabetes. As a result, this can decrease the prevalence of the disease and improve initial management practices. Furthermore, this will increase patients' satisfaction and overall quality of care.

A comprehensive diagnostic framework has the potential to streamline medical services and empower patients. Machine learning based on algorithms offers a unique tool for healthcare professionals to utilize, from both an epidemiological and treatment perspective. From a systems standpoint, the ability to centralize medical data and predict trends in population health would allow resource allocation to the identified gaps, which in turn, strengthens the population's health.

Moreover, another meaningful impact of integrating machine learning is the benefit to the patients. The WHO [31] defines empowerment as "a process through which people gain greater control over decisions and actions affecting their health" and affects both individual and community levels. To empower the population, they must get access to their information and have it delivered in...
an understandable format, transparent, and overall—user-friendly. Ease of use is paramount for patient engagement.

Literature indicates that within the 21st century, mobile health technologies resulted in increases in connecting users on a community, population, and global level [32]. Mobile health addresses the rising burden of chronic diseases while encouraging health systems to shift towards patient-centric designs [32]. Mobile health consists of medical practice supported by a portable diagnostic device [32]. The use of these devices at the point-of-care resulted in not only a change in healthcare delivery, but an increase in patient engagement, a reduction in healthcare costs, and improved patient prognosis [32].

Model learning has the potential to increase patient empowerment via mobile health. The compatibility for our connected world through accessibility from a smartphone, desktop or other personal electronic devices, in the way of an app, is potentially highly useful in capitalizing on our mobile interconnectedness. However, before the implementation of mobile health, guidelines to manage these machine learning models are essential for healthcare.

Systematically developed statements based on research, best practices, best scientific evidence, and experience act as guidelines [31,33]. Guidelines support healthcare providers with an outline for patient care to ensure that individuals receive the same or similar patient care across healthcare facilities.

Standardizing guidelines across healthcare facilities can aide authorities in bridging the gap between research and practices within these facilities, which helps to foster consistent services. Additionally, standardizing guidelines across healthcare facilities helps healthcare providers to identify the what, where, when, and how of the patient's health; while collecting, sharing, and reporting data improves and streamlines the process. The collected and reported data based on clinical guidelines assists healthcare and public health authorities in identifying the age groups or individual patients at high risk of having diabetes (Type 1 or Type 2).

The collected and reported data assists healthcare authorities in planning prevention and treatment plans. The flexibility of the process allows healthcare authorities to navigate the ever-changing healthcare landscape. Gaps, limitations, and needs of population health are dynamic, and to avoid steep healthcare costs, the allocation of resources must have a basis in evidence to resolve pressing issues best.

**Conclusion:**

In this paper, the building of a machine learning model for early prediction of diabetes had a basis on real health data collected from the Ministry of National Guard Health Affairs, Saudi Arabia. The comparison of four machine learning algorithms, namely Deepnet, models (decision tree), ensemble and Logistic Regression, used 17 attributes. Under assessment, Deepnet achieves the best result using the four different evaluation criteria. This paper demonstrates that machine learning-based algorithms have excellent potential in predicting diabetes with high accuracy. Future work is to evaluate the model on a larger data-set and use the model with the Internet of Things devices.
Acknowledgement:

This study was funded by the King Abdullah International Medical Research Center (KAIMRC), National Guard, Health Affairs, Riyadh, Saudi Arabia, with research grant No. RC17/248/R.

References


Improving Accuracy for Diabetes Mellitus Prediction by Using Deepnet


