USING DEEP LEARNING ON EHR DATA TO PREDICT DIABETES

by

THOMAS GARSKE

B.S., Electrical Engineering, University of St. Thomas, 2010

B.A., Physics, University of St. Thomas, 2010

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Master of Science

Electrical Engineering

2018
This thesis for the Master of Science degree by

Thomas Garske

has been approved for the

Electrical Engineering Program

by

Dan Connors, Advisor

Chao Liu

Tim Lei

September 10, 2018
ABSTRACT

Type 2 Diabetes Mellitus (T2DM) is a leading cause of death and disability in the United States with high prevalence, growing incidence, and serious adverse outcomes. A major challenge of addressing chronic health conditions such as T2DM is that often diagnosis and treatment are not established until after the disease has progressed and patients are already suffering. Traditional strategies to predict future risk of diabetes have generally used demographic and clinical data from prospective cohort studies and statistical models and risk scores. Such strategies appear to perform moderately well, but, in general, have limited success when applied to different patient populations [13].

Advancements in machine learning in medicine have recently made headlines in the area of disease prediction and care prescription. Modeling and predictive techniques work best when varieties of health data can be compiled and trained on. The generation of electronic record keeping has greatly impacted the ability for new machine learning techniques to be applied to healthcare data. This ability was accelerated by the Affordable Care Act, which incentivized the proper use and adoption of electronic health records (EHRs) [17]. Combining machine learning techniques with EHR data provides an unique opportunity to improve care management at the population level by predicting disease onset and progression. When care management can assess a patient’s current status, predict future trends, and assign the most prescriptive actions, patient outcomes can be improved while simultaneously lowering healthcare costs.

Deep Learning is a subfield of the broader field of machine learning. Most deep learning methods learn through an artificial neural network, otherwise known as a Deep Neural Network (DNN). Neural networks are made up of several layers of connected processors called neurons. In deep learning, each layer transforms its input into a slightly more abstract
representation and is about accurately assigning credit across many such layers [28]. Other
deep learning models to forecast diabetes onset in future years were largely limited by small
population sizes and had an ineffective level such as 73% accuracy [24].

The objective of this thesis is to create a diabetes machine learning dataset from EHR
data, and develop an optimized deep learning model to identify patients at risk of receiving
a new diagnosis of T2DM during the following 24 month time period for a general popula-
tion. A diabetes prediction model was successfully developed using a DNNClassifier from
TensorFlow and optimized by developing a framework for parameter search over different
optimization functions, activation functions, sampling methods, and scaling methods. The
resulting diabetes machine learning dataset contains a feature set of 22 attributes for 149,050
patients and related class labels for a diabetes diagnosis within 6, 12, 18, and 24 month time
periods. The best model for prediction of diabetes had 83% true positive and 88% true
negative accuracy scores for prediction within 24 months. Early identification of the risk of
T2DM in these patients provide opportunities for care intervention to promote better long
term outcomes.

The form and content of this abstract are approved. I recommend its publication.

Approved: Dan Connors
# TABLE OF CONTENTS

## CHAPTER

I. **INTRODUCTION** ................................................................. 1

II. **BACKGROUND** .............................................................. 4

   Type 2 Diabetes Mellitus ...................................................... 4
   T2DM Complications and Risk Factors ..................................... 4
   Machine Learning .............................................................. 5
   Supervised, Semi-supervised, and Unsupervised Learning .......... 6
   Deep Learning ................................................................. 7
   Prediction Model Optimization Parameters ............................. 7
   Optimization Functions ....................................................... 7
   Activation Functions ......................................................... 8
   Sampling Methods ............................................................. 8
   Scaling Methods .............................................................. 9

   Machine Learning Dataset .................................................. 9

III. **APPROACH** ................................................................. 11

   Running the experiments ................................................... 13

IV. **EXPERIMENTAL RESULTS** ............................................... 14

   Machine Learning Diabetes Dataset ..................................... 14
   Diabetes Prediction Model .................................................. 17
   Model Validation ............................................................. 19
   Model Feature Breakdown ................................................... 19

   Model Optimization Parameter Search ................................ 23

V. **CONCLUSION** ............................................................... 29

REFERENCES ............................................................................. 32
CHAPTER I

INTRODUCTION

Type 2 Diabetes Mellitus (T2DM) is a serious public health issue in the United States. Incidence of T2DM is growing at an unprecedented rate, and previous estimates to predict the growth of the disease have drastically fallen short. In 2005, it was estimated that the global number of adults with T2DM would grow to 380 million by 2025 [15]; however, that prediction had been surpassed before 2015 [8]. It is now estimated that there are 425 million adults living with diabetes globally. The U.S. prevalence rates for diabetes among adults who are 20 years or older is now over 30 million (10.8%), with an additional estimated 11.5 million adults with undiagnosed diabetes [9]. Diabetes is the sixth overall leading cause of death in the United States with nearly 80,000 deaths in 2015 [5]. Often, T2DM is asymptomatic in early stages of the disease, and can remain undiagnosed for years. Diabetes management is most successful in reducing the burden of diabetes and related complications through early diagnosis and intervention [2].

Advancements in machine learning in medicine have recently made headlines in the area of disease prediction and care prescription. Modeling and predictive techniques work best when varieties of health data can be compiled and trained on. The generation of electronic record keeping has greatly impacted the ability for new machine learning techniques to be applied to healthcare data. This ability was accelerated by the Affordable Care Act, which incentivized the proper use and adoption of electronic health records (EHRs) [17]. The medical records in EHRs are often extensive and contain detailed clinical information from outpatient visits, including data for allergies, appointments, medications, lab results, patient demographics, diagnoses, vaccines, and vitals. Machine learning tools and their optimized deployment are a valuable tool in analyzing large datasets and can be applied to patient EHR data; providing an unique opportunity to improve care management at the population level through insight into disease onset and progression within various patient cohorts. When care management can assess a patient’s current status, predict future trends, and assign the
most prescriptive actions, patient outcomes can be improved while simultaneously lowering healthcare costs.

Traditional strategies to predict future risk of diabetes have generally used demographic and clinical data from prospective cohort studies, and statistical models or risk scores. In general, such strategies appear to perform well but have limited success when applied to different patient populations [13]. Prior diabetes prediction studies have demonstrated feasibility in using EHR data for prediction of T2DM in patients, but the results were limited by the number of patients available for training [19]. With access to Allscripts’ de-identified data lake containing over 50 million patient lives, this research has the unique opportunity to create a diabetes machine learning dataset with a large and diverse patient cohort.

The work of this thesis was to evaluate an EHR based deep learning model to predict T2DM diagnosis in a general population. The objective is to deliver a diabetes risk forecasting model using EHRs for prediction of T2DM to allow early intervention treatment. This work expands diabetes prediction models by increasing the number of patients and utilizing a much broader patient dataset than was previously available to researchers. The contribution of that work includes:

1. Developing a validated deep learning model to predict the timeline of a new diagnosis of T2DM within the following 6, 12, 18, and 24 month time periods from 12 months of a patient’s clinical EHR history. The model was developed by training a deep learning artificial neural network using statistical learning on features selected through domain knowledge of common risk factors for T2DM.

2. Building a framework for development and evaluation of optimized classifiers using a parameter search over selected optimization functions, activation functions, scaling methods, and sampling methods. This framework was made to be general purpose and can be used on any deep neural net binary classification problem. After reading this thesis you will have an understanding of some insights into parameters provided by this framework. For example, for the dataset generated in this research, the effects
of dataset imbalance are apparent when comparing sampling methods. Additionally, some optimization functions are more likely to get stuck on an incorrect solution.

3. Generating a diabetes machine learning dataset that includes relevant EHR data across a general distribution of the United States. This expands the potential usability of this prediction model as a clinical tool by increasing the number of patients and utilizing a much broader patient dataset than was previously available to researchers.

This thesis is organized as follows: Chapter II covers the background of T2DM, EHR data, development of machine learning models, and the optimization of those models. Chapter III examines the approach to the diabetes machine learning dataset creation, prediction model development, and model parameter optimization. The experimental results section, Chapter IV, shows details and specifics on the resulting machine learning dataset, performance data for the prediction of various diagnosis timelines, validation of those prediction models, and the results of the model parameter optimization search. Finally, Chapter V will summarize these results and discuss potential further experiments.
CHAPTER II

BACKGROUND

Type 2 Diabetes Mellitus

Diabetes is a disease that impairs the body’s ability to respond to insulin, a hormone needed to allow glucose to enter cells, resulting in higher levels of glucose in the blood. While Type 2 Diabetes Mellitus (T2DM) is the most common form of diabetes, there are other chronic diabetic conditions including prediabetes, Type 1 Diabetes Mellitus (T1DM), and gestational diabetes [22]. In prediabetes and T2DM, the pancreas is increasingly unable to produce enough insulin as cells gradually become resistant to the affects of insulin. T1DM is a condition in which the body cannot effectively regulate blood sugar as the pancreas produces little to no insulin. While the cause is unknown, T1DM is believed to be genetic. Some women develop gestational diabetes during pregnancy, but it usually resolves after delivery. Individuals that have had gestational diabetes are at higher risk of developing T2DM, and sometimes gestational diabetes diagnosed during pregnancy is actually T2DM [4].

T2DM Complications and Risk Factors

Long-term complications of T2DM develop gradually, but those complications can be disabling or life-threatening. Potential major complications include cardiovascular disease, neuropathy, nephropathy, retinopathy, foot or leg amputation, bacterial and fungal infections, hearing impairment, dementia, and depression. The risk of complications increases with poor blood sugar control and the length of time with T2DM [21]. It is not fully understood what makes some individuals more susceptible to developing T2DM. However, evidence indicates that certain factors increase the risk of developing the disease including:

- Obesity. High amounts of body fat can cause the endoplasmic reticulum (ER), a system of cellular membranes, to suppress the signals of insulin receptors and lead to insulin resistance in cells [23].

- Family history. Individuals with first-degree relatives who have T2DM are at approximately 3.5 times as much risk of T2DM compared to the general population [25].
Multiple large scale population studies have demonstrated the existence of genetic susceptibility variants linked to T2DM risk [31, 27, 33].

- Race. Certain racial and ethnic backgrounds have been shown to contribute to variations in glycemic control and overall risk of developing T2DM for both adult and youth populations [11, 26].

- Age. While incidence of T2DM in youth is rising at an alarming rate, diabetes is more prevalent in adult and aging populations. Much of this can be attributed to compounding years of poor diet and insufficient physical activity leading to chronic insulin resistance [12].

- Gender. The risk of developing prediabetes and T2DM is increased for women that were previously diagnosed with gestational diabetes [4]. Women that have given birth to babies weighing outside of the normal weight distribution (both low weight and high weight) exhibit a higher risk of developing T2DM [10]. Having polycystic ovary syndrome increases the risk of T2DM [18].

- High blood pressure. High blood pressure readings of over 140/90 millimeters of mercury (mm Hg) is highly common among individuals with T2DM.

- Abnormal cholesterol. Abnormal cholesterol and triglyceride levels often accompanies T2DM. Risk increases with high levels of triglycerides and low levels of high-density lipoprotein (HDL) [3].

**Machine Learning**

Machine learning is the process of equipping computers with the ability to learn without explicit programming through computer programs that can change or adapt when exposed to new data. There are four general categories of machine learning.

- Classification, predicting a category or label. Classification involves training on observations to create a model that will attempt to predict whether an observation belongs
to a given category. For example, classifying an image as a “cat” or “dog”. A model would train on labeled images of dogs and cats, and predict the similarity to a dog or a cat for a previously unseen photo.

- Regression, predicting a value. Regression problems try to predict the value of a continuous variable based on previous information. Regression is similar to classification, however, the goal is to find a real value rather than the category of an observation.

- Clustering, finding the hidden structure of data. Clustering is a learning method that attempts to predict where groups of similar observations are located. This method attempts to group a set of objects by similarity so that objects in one group are more similar to each other than to objects in another group.

- Dimensionality reduction, obtaining a set of principal variables. Dimensionality reduction is the process of reducing dimensionality by obtaining a set of principal variables. Removing data that is dis-informative can allow a model to find more general classification regions and yield better performance [32].

Each of these categories can be described by three fundamental components: representation, which is how the data is structured; evaluation, or scoring of hypotheses; and optimization, which is the process of generating models. All machine learning methods are a combination of these elements, and these components represent the framework to understand any machine learning algorithm [14].

Supervised, Semi-supervised, and Unsupervised Learning

In supervised learning, it is important that labels are available to the algorithm during the training of the function so the labels of actual values can be compared to the predicted labels to evaluate how the trained model is performing. Labeling is often a time intensive and manual process.

Techniques that do not require labeled data, such as clustering, are known as unsupervised learning methods. Evaluating the performance of unsupervised learning models is
more complicated than supervised learning models because there are no true labels to test against.

Semi-supervised learning techniques overlap between supervised and unsupervised learning methods and could include both labeled and unlabeled observations. For example, clustering can be used to group the data by similarity. Labels can then be assigned to the unlabeled observations by combining information about the labels and information about the clusters. This process would provide more labeled information for a supervised learning method [29].

**Deep Learning**

Deep Learning is a subfield of the broader field of machine learning in which most methods learn through a Deep Neural Network (DNN). Neural networks are made up of several layers of connected processors called neurons. The output of an active neuron is determined by its activation function. Input neurons are activated by the input feature set, while hidden layer neurons are activated through weighted connections to previously active neurons. In deep learning, each layer transforms its input into a more abstract representation and the goal of the neural net is to accurately transform the data across many such layers [28].

TensorFlow is an open source library developed by researchers from the Google Brain team for high performance computation. TensorFlow provides support for both machine learning and deep learning, and is used across many scientific domains [1]. The model in this study was built using TensorFlow’s Deep Neural Network Classifier (DNNClassifier), a pre-built deep learning estimator.

**Prediction Model Optimization Parameters**

**Optimization Functions**

Optimization algorithms are what the model uses to minimize or maximize an Error function: a function used in computing the target predictions. This study evaluated the following optimization functions as a part of the DNNClassifier parameter optimization search:
Adaptive optimizer functions can change learning rate, a hyper-parameter that controls how much to adjust the weights of the network with respect to the loss gradient, while training. Momentum allows the optimization function to remember the change of the previous step for each iteration, and produce the next update through a combination of the gradient and that last change. Proximal optimizer functions use a generalized form of projection to consider proximity of the features when adjusting weights.

**Activation Functions**

In neural networks, the activation function, sometimes referred to as the transfer function, of a node returns the output value of that node given a set of inputs. This function determines if the output of a node will be a 1 or a 0. Activation functions evaluated in this thesis: Rectified Linear Unit (Relu), the default setting for the DNNClassifier and most commonly used function; Exponential Linear Unit (elu); Scaled Exponential Linear Unit (Selu); Softplus, a smooth approximation of Relu; tanh; and softsign, a smooth approximation of tanh.

**Sampling Methods**

Oversampling and undersampling are roughly equivalent and opposite techniques in data analysis that adjust the distribution of classes for a dataset. Combined samplers perform a combination of under- and over-sampling offering the best performance, but take a longer time to process the data. Scikit-learn sampling methods evaluated in this project are: RandomOverSampler, which randomly duplicates members from the under-represented class; RandomUnderSampler, which randomly drops members from the over-represented class; and Synthetic Minority Over-sampling Technique (SMOTE), which creates new records that represents minority groups in the under-represented class and attempts to fill in gaps [6].
Figure II.1 shows how the different sampling methods modify an example dataset.

**Scaling Methods**

Scalers are used to normalize the data to reduce size and improve performance. The scaling methods used in the parameter optimization are: MinMaxScaler, normalizes the data from -1 to 1; StandardScaler, normalizes the data and centers on the calculated mean from -1 to 1; RobustScaler, does the same as StandardScaler but eliminates the bottom and top quartiles to drop outliers; and MaxAbsScaler, which scales each feature by its maximum absolute value.

**Machine Learning Dataset**

In healthcare, an Electronic Health Record (EHR) is broadly defined as longitudinal data that represents the full health record of medical services provided to a patient [7]. As EHRs are gaining traction and the applications become more advanced, the quality of the data improves. For this thesis, the source EHR system for the dataset was a combination of Allscripts Touchworks EHR, a cloud-based solution for midsize to large practices, and
Allscripts Professional EHR, a solution for small to midsize practices. Both of these EHRs are designed for the Outpatient (a patient who receives medical treatment without being admitted to a hospital) setting, and does not contain Inpatient (a patient who stays in a hospital while under treatment) information.

The Allscripts clinical data warehouse contains de-identified Protected Health Information (PHI) for over 50 million patients over a 10-year period. Table II.1 shows the tables information available in the Allscripts clinical data warehouse.

Table II.1: Allscripts’ clinical data warehouse tables.

<table>
<thead>
<tr>
<th>Table Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergies</td>
<td>Allergy information per patient</td>
</tr>
<tr>
<td>Appointments</td>
<td>Patient appointments scheduled and attended</td>
</tr>
<tr>
<td>Encounters</td>
<td>Anytime a record is updated or accessed</td>
</tr>
<tr>
<td>Medications</td>
<td>Medication prescriptions and related problems (e.g. back pain)</td>
</tr>
<tr>
<td>Orders</td>
<td>Orders submitted for lab results</td>
</tr>
<tr>
<td>Patients</td>
<td>Demographic information per patient</td>
</tr>
<tr>
<td>Problems</td>
<td>Diagnosis codes (e.g. ICD9 and ICD10 codes for diabetes)</td>
</tr>
<tr>
<td>Providers</td>
<td>Healthcare provider (physicians and hospitals) information</td>
</tr>
<tr>
<td>Results</td>
<td>Results for labs (HbA1c, cholesterol levels, blood work, etc.)</td>
</tr>
<tr>
<td>Vaccines</td>
<td>Vaccine information for patients</td>
</tr>
<tr>
<td>Vitals</td>
<td>Vitals records for patients (blood pressure, BMI, etc.)</td>
</tr>
</tbody>
</table>
CHAPTER III
APPROACH

The python and SQL scripts to build out the diabetes machine learning dataset, run
the parameter search experiments, and validate the diabetes prediction models are in a
Jupyter Notebook: an interactive document format for publishing code, results, and expla-
nations [16]. This experiment requires the creation of a local virtual environment using a
bash setup script to create the local environment, install the pip modules, and start the
Jupyter notebook server. The first step of the Jupyter notebook includes the import of
relevant python modules and iPython extensions sets the experimental parameters.

The first step in building a classifier is to perform feature engineering, the process of using
algorithms or domain knowledge to create features, on the dataset. Deep Neural Networks
(DNN) can only operate on float values as every node can perform multiplication or addition
operations. As a result, even if the source data is categorical by nature, a machine learning
model represents all features as numbers. If a parameter is categorical, it can be converted to
indicator column features through generation of dummy variables. Continuous parameters
can be normalized and scaled, or simply binned to create indicator columns.

Figure III.1 shows process for creating the machine learning dataset and building a
diabetes prediction classifier. To create the machine learning dataset, the code connects to
the Allscripts’ SQL Data Warehouse using the latest Open Database Connectivity (ODBC)
driver for SQL Server. Tables containing the serialized data were built for each of the selected
features from patient demographics, problem diagnoses, vitals and lab results tables. This
data was aggregated and joined into a flat (1 row per patient) table and exported to an
external file using a pyodbc connector. A data preparation class loads the data file, excludes
any unknown or null values, and outputs X and Y tensors and the resulting datasets.

DNNClassifiers expect data to be in X and Y tensors of floats; where X represents
the feature set of predictors and Y represents the target class information to be predicted.
After it has been split into X and Y tensors, the data is further divided into Train, Test
Figure III.1: Building a diabetes machine learning dataset and classifier.

and Validation datasets. A common approach is to use 70% of the dataset for training and reserve the remainder for testing and validation. It is critical to split the dataset before fitting scalers, samplers, or classifiers to ensure that the model does not train on any calculated values of mean or standard deviation from the test data. At this stage the dataset is sampled to address any imbalanced class imbalance, and scaled to improve the performance of training the model.

The DNNClassifier is built using the selected model parameters, and trained on the training dataset. A validation monitor looks to maximize the precision recall value and minimize the loss during training by evaluating model performance on the validation dataset. Once those values stop changing significantly the monitor stops the model from training further. This validation monitor is critical in maximizing the efficiency of the parameter optimization search framework. The model is then evaluated using the testing dataset.
comparing the predicted labels to the actual labels through a variety of methods including accuracy, precision-recall, and the confusion matrix.

Running the experiments

A helper class was developed to iterate through all selected sampling methods, scaling methods, optimizer functions, and then activation functions. A function builds, trains, and evaluates a diabetes prediction model for each configuration. As each model is evaluated, a function captures the resulting accuracy and performance metrics and appends the results to a Pandas dataframe in python. The dataframe is serialized to a text file as each result is appended. This setup allows the experiments to be paused and resumed; a beneficial feature as it can take days to complete.

These results are plotted using matplotlib to display relevant trends and to allow conclusions to be drawn from the different aspects of the parameter optimization search. After the best inputs have been found, the script uses the suggested parameters to train and optimize a diabetes prediction model. This selected model is evaluated further by comparing the results with a traditional diabetes risk score method.
CHAPTER IV
EXPERIMENTAL RESULTS

Machine Learning Diabetes Dataset

Using the Allscripts’ Clinical Data Warehouse with EHR history of over 50 million patients, this project was successful in curating a diabetes prediction machine learning dataset consisting of 22 attributes from encounters gathered from June 2015 through June 2016, and diabetes diagnosis target values from June 2016 through June 2018. The target value, a future diagnosis of diabetes, was determined by prevalence of International Classification of Diseases, Ninth Revision (ICD-9) and International Classification of Diseases, Tenth Revision (ICD-10) codes. While ICD-10 codes were implemented in October 2015, the codes were not immediately adopted universally, so any study relying on these codes needs to search for diagnosis in both of the code-sets [30].

Table IV.1 shows the SQL tables in the data warehouse and which features were extracted from each. Any patient that had a prior diagnosis of T2DM, gestational diabetes, or personal history of metabolic disease before June 2016 was excluded. Patients that did not have at least one active encounters in all of the prediction time frames, and at least one new problem diagnosis in any of the time frames, were excluded due to inactivity in the EHR system. The diabetes machine learning dataset that was created contains features for 149,050 patients. Figure IV.1 shows the diabetes and non-diabetes imbalance for each of the targets for those patients.
Table IV.1: Diabetes machine learning dataset dimensions.

<table>
<thead>
<tr>
<th>Table</th>
<th>Dimension</th>
<th>Feature Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Demographics</td>
<td>Age</td>
<td>Continuous</td>
</tr>
<tr>
<td>Patient Demographics</td>
<td>Gender: Male</td>
<td>Indicator</td>
</tr>
<tr>
<td>Patient Demographics</td>
<td>Race: White</td>
<td>Indicator</td>
</tr>
<tr>
<td>Patient Demographics</td>
<td>Ethnicity: Hispanic</td>
<td>Indicator</td>
</tr>
<tr>
<td>Problem Diagnoses</td>
<td>Abnormal Blood Pressure</td>
<td>Indicator</td>
</tr>
<tr>
<td>Problem Diagnoses</td>
<td>Family History of Diabetes</td>
<td>Indicator</td>
</tr>
<tr>
<td>Lab Results</td>
<td>Minimum Hemoglobin A1c (HbA1c)</td>
<td>Continuous</td>
</tr>
<tr>
<td>Lab Results</td>
<td>Maximum Hemoglobin A1c (HbA1c)</td>
<td>Continuous</td>
</tr>
<tr>
<td>Lab Results</td>
<td>Average Hemoglobin A1c (HbA1c)</td>
<td>Continuous</td>
</tr>
<tr>
<td>Lab Results</td>
<td>Most Recent Hemoglobin A1c (HbA1c)</td>
<td>Continuous</td>
</tr>
<tr>
<td>Vitals</td>
<td>Minimum Body Mass Index (BMI)</td>
<td>Continuous</td>
</tr>
<tr>
<td>Vitals</td>
<td>Maximum Body Mass Index (BMI)</td>
<td>Continuous</td>
</tr>
<tr>
<td>Vitals</td>
<td>Average Body Mass Index (BMI)</td>
<td>Continuous</td>
</tr>
<tr>
<td>Vitals</td>
<td>Most Recent Body Mass Index (BMI)</td>
<td>Continuous</td>
</tr>
<tr>
<td>Vitals</td>
<td>Minimum Diastolic Blood Pressure</td>
<td>Continuous</td>
</tr>
<tr>
<td>Vitals</td>
<td>Maximum Diastolic Blood Pressure</td>
<td>Continuous</td>
</tr>
<tr>
<td>Vitals</td>
<td>Average Diastolic Blood Pressure</td>
<td>Continuous</td>
</tr>
<tr>
<td>Vitals</td>
<td>Most Recent Diastolic Blood Pressure</td>
<td>Continuous</td>
</tr>
<tr>
<td>Vitals</td>
<td>Minimum Systolic Blood Pressure</td>
<td>Continuous</td>
</tr>
<tr>
<td>Vitals</td>
<td>Maximum Systolic Blood Pressure</td>
<td>Continuous</td>
</tr>
<tr>
<td>Vitals</td>
<td>Average Systolic Blood Pressure</td>
<td>Continuous</td>
</tr>
<tr>
<td>Vitals</td>
<td>Most Recent Systolic Blood Pressure</td>
<td>Continuous</td>
</tr>
</tbody>
</table>
Figure IV.1: Diabetes target label distribution.
Diabetes Prediction Model

Accuracy is the overall percent of predictions that are correct, and while useful, can be misleading when the dataset is imbalanced. Precision, recall, and confusion matrices are all methods that provide insight into model performance for both positive and negative classification and accuracies. In binary classification, precision is the fraction of positive value subjects among the evaluated subjects. Recall is the fraction of positive value subjects that have been evaluated over the total amount of positive value subjects. The precision and recall curve of a model provides insight into how well the model performed in predicting a positive value (precision) over the number of test subjects that had a positive value (recall). Figure IV.2 demonstrates how to read a confusion matrix and the definitions of a True Positive (TP), a True Negative (TN), a False Positive (FP), and a False Negative (FN) when comparing predicted and true labels.

The best model for predicting diabetes would be the model with the best precision, and the highest TP and TN accuracies. Figure IV.3 shows the normalized confusion matrices for predicting diabetes within 6, 12, 18, and 24 months. These graphs show that the model performs reasonably well on predicting onset of diabetes for all targets, and the longer time-frames have increasing improvement on true-negative prediction of the absence of diabetes.
Figure IV.3: Confusion matrices for predicting diabetes within target time-frames.

Table IV.2 shows the optimized configuration and the performance metrics of the diabetes prediction model for each of the target time-frames. In this research, the best model for each target was developed using a different combination of parameters.

Table IV.2: Model configurations for best model results.

<table>
<thead>
<tr>
<th>Target</th>
<th>Model Configuration</th>
<th>Prec.</th>
<th>Acc.</th>
<th>TP</th>
<th>TN</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 months</td>
<td>softsign, graddesc, naiveover, robust</td>
<td>10%</td>
<td>65%</td>
<td>80%</td>
<td>64%</td>
</tr>
<tr>
<td>12 months</td>
<td>elu, momentum, naiveover, robust</td>
<td>23%</td>
<td>66%</td>
<td>82%</td>
<td>64%</td>
</tr>
<tr>
<td>18 months</td>
<td>softplus, proxadagrad, smote, standard</td>
<td>57%</td>
<td>74%</td>
<td>84%</td>
<td>70%</td>
</tr>
<tr>
<td>24 months</td>
<td>elu, momentum, smote, robust</td>
<td>91%</td>
<td>86%</td>
<td>83%</td>
<td>88%</td>
</tr>
</tbody>
</table>
Model Validation

The deep learning diabetes prediction model outperformed traditional methods. This model was validated by comparing results to the American Diabetes Association Type 2 Diabetes Risk Test [20]. The ADA Risk Test consists of seven metrics based on age, gender, gestational diabetes, family history, physical activity, and BMI. The research was able to produce a score using 5 of the 7 metrics. Gestational diabetes was excluded due to low diagnosis rates (<0.1%), and physical inactivity is not available in EHR datasets. A score of 5 or higher from the above metrics is a ‘high risk’ - considered a ‘prediction’ of diabetes diagnosis for the purpose of validating this model. Figure IV.4 shows the normalized confusion matrices for predicting diabetes within 6, 12, 18, and 24 months compared with the results of the American Diabetes Risk Test. The deep learning diabetes prediction outperformed the risk test in accurately predicting both true positives and true negatives for all time frames.

Model Feature Breakdown

One of the criticisms of using deep learning prediction models is the lack of transparency into how the model is fitting to the features. In healthcare, physicians need this insight to effectively implement any machine learning model in a clinical setting. As all of the models were trained on the same patient dataset, the 24 month prediction model was selected for further analysis as it had demonstrated the best accuracy and precision in predicting diabetes. Figure IV.5 shows the breakdown for each indicator column for correct and incorrect predictions. Broadly, these graphs show that the model is fitting to the dataset well. They also provide insight into which features may require further exploration, or development of new models specifically tailored to those features, for improved specificity. Figure IV.6 shows the breakdown for each continuous column for correct and incorrect predictions. The model performed better on the higher and lower ranges for HbA1c values, and on higher diastolic and systolic blood pressure values.
Figure IV.4: Confusion matrices for predicting diabetes within target time-frames.
Figure IV.5: Indicator column breakdown for correct and incorrect predictions.
Figure IV.6: Continuous column distribution for correct and incorrect predictions.
Model Optimization Parameter Search

The final goal of this thesis is the development of a model parameter optimization framework to identify the optimization function, activation function, scaling method, and sampling method for the best model performance. The framework developed as a result of this research provided a strong basis for model optimization by building, training, and evaluating 768 configurations of the DNNClassifier for each of the 4 diabetes target time-frames. The output of this framework is a list of configurations to explore and further improve through learning rate tuning and hidden layer adjustment. While training, each model monitors the loss function and evaluates model accuracy; stopping training when the model is sufficiently fitted to the data. This validation monitor allows the framework to minimize time spent training each model.

Figure IV.7 and figure IV.8 show the distribution of accuracy and precision for all configurations generated by the model parameter optimization search for each of the target time-frames. This figure demonstrates how dataset imbalance can affect training. As the target classes become less imbalanced, training is less likely to lock onto one class and classify all observations as diabetic or non-diabetic.

Figure IV.9 shows the results of the scaling method model optimization. This analysis is useful in that it shows that the scaling method, while critical, does not negatively impact the training results.

Figure IV.10 shows the results of the sampling method model optimization. As the 6 and 12 month target classes are highly imbalanced, this figure shows that selecting a sampling method is a critical decision for a successful model. Interestingly, SMOTE sampling performed best on the least imbalanced dataset, and under performed compared to naive methods on strongly imbalanced targets.

Figure IV.11 shows the results of the optimizer function model optimization. For this dataset, RMS Propagation and the Adaptive Momentum optimizers were likely to over adjust
Figure IV.7: Range of results for each model over the parameter optimization search.

Figure IV.8: All results for each model over the parameter optimization search.
and lock onto one of the target classes for all predictions. This insight is useful in successfully training models for this dataset because it allows the researcher to avoid optimizer functions that do not properly fit to the feature set.

Figure IV.12 shows the results of the activation function model optimization. While there is minor variance between methods, activation function had the least impact on model performance.
Figure IV.10: Effect of sampling method parameter on model performance.
Figure IV.11: Effect of optimizer function parameter on model performance.
Figure IV.12: Effect of activation function parameter on model performance.
CHAPTER V
CONCLUSION

The work of this thesis was to evaluate an EHR based deep learning model to predict T2DM diagnosis in a general population. The objective is to deliver a diabetes risk forecasting model using EHRs for early prediction of T2DM to allow for early intervention and treatment. The contribution of that work includes the creation of an EHR based machine learning dataset, the development of a deep learning diabetes prediction model, and the development of a framework for building optimized classifiers.

The diabetes machine learning dataset created represents EHR data across a general distribution of the population of the United States. This expands the potential usability of this prediction model as a clinical tool by increasing the number of patients and utilizing a much broader patient dataset than was previously available to researchers. The use of a large clinical dataset to create machine learning datasets provides significant value because of the vast availability of datasets for various populations. There is opportunity for future research to use the same EHR dataset for prediction of other chronic diseases. Further investigation into T2DM prediction should evaluate the feature set creation. While there is over a decade of EHR data available, this study was restricted to 12 months of clinical history per patient. Future studies should explore expanding the feature set to include multiple years of clinical history. Additionally, this study relied on ICD9 and ICD10 diagnosis codes for the target labels of diabetes diagnosis. These diagnosis codes are potentially an unreliable label and future research should validate diabetes diagnosis labels through other means such as HbA1c scores or the presence of diabetes medication prescriptions.

The next contribution was to develop and evaluate a deep learning model to predict the timeline of a new diagnosis of T2DM within the following 6, 12, 18, and 24 month time periods from 12 months of a patient's clinical EHR history. The model was developed through statistical learning and using features selected through domain knowledge of common risk factors for developing T2DM. Validation of this model (see figure IV.4) demonstrates that the
diabetes prediction model developed through this research outperforms the ADA Diabetes Risk Test when applied to a general EHR population. The feature exploration of the diabetes prediction model provides insight into how further research might improve accuracy through various feature engineering methods. Based on the results of figure IV.6, future research may consider binning some of the continuous feature columns to provide improved context to the model. For example, binning BMI values into ‘normal,’ ‘overweight,’ or ‘obese’ categories could function to improve model accuracy. Similarly, Figure IV.5 shows that improved context could be achieved by applying separate models to separate feature categories for improved specificity of the dataset.

The work of this thesis included development of a framework for developing optimized classifiers through a parameter search over selected optimization functions, activation functions, scaling methods, and sampling methods. This framework was made to be general purpose and can be used on any deep neural net binary classification problem. As demonstrated in this research, the model optimization framework provides insight into how different model parameters performed. For example, for the dataset generated in this research, the effects of dataset imbalance are apparent when comparing sampling methods. Figure IV.10 shows that, as the 6, and 12 month targets are highly imbalanced, selecting a sampling method is critical for a successful model. Additionally, Figure IV.11 shows that some optimization functions are more likely to get stuck on an incorrect solution. The review of the model optimization framework demonstrates a good foundation to assist future endeavors to develop models for other disease cohorts. Future work to expand the framework should consider including a grid search over the hidden units of the neural network. This search should be either a growth method of starting small and building out or a paring method of starting with a large network and reducing. If the goal is performance, the former should be used, if the goal is to prevent over-training, the latter.

The outcome of building a deep learning model on EHR data demonstrates that there is an application for machine learning in healthcare, and the feasibility of a models potential
effectiveness as a prediction tool in a clinical setting. Anything that can be done to identify dangerous diseases early in patients and provide opportunities for care intervention is valuable for population health management and improving healthcare. Early identification of the risk of T2DM in these patients provides opportunities for care intervention to promote better long-term outcomes. With the continued widespread adoption of electronic health records, deep learning prediction models will undoubtedly continue to be a significant area of advancement in healthcare.
REFERENCES


