**APS360 Winter 2020 Project Proposal**

**Team 62**

**Sean Llera - 1004928536**

**Nathan Jones - 1005003023**

**Priscilla Deng - 1005227826**

**Sadra Tonekaboni - 1004899461**

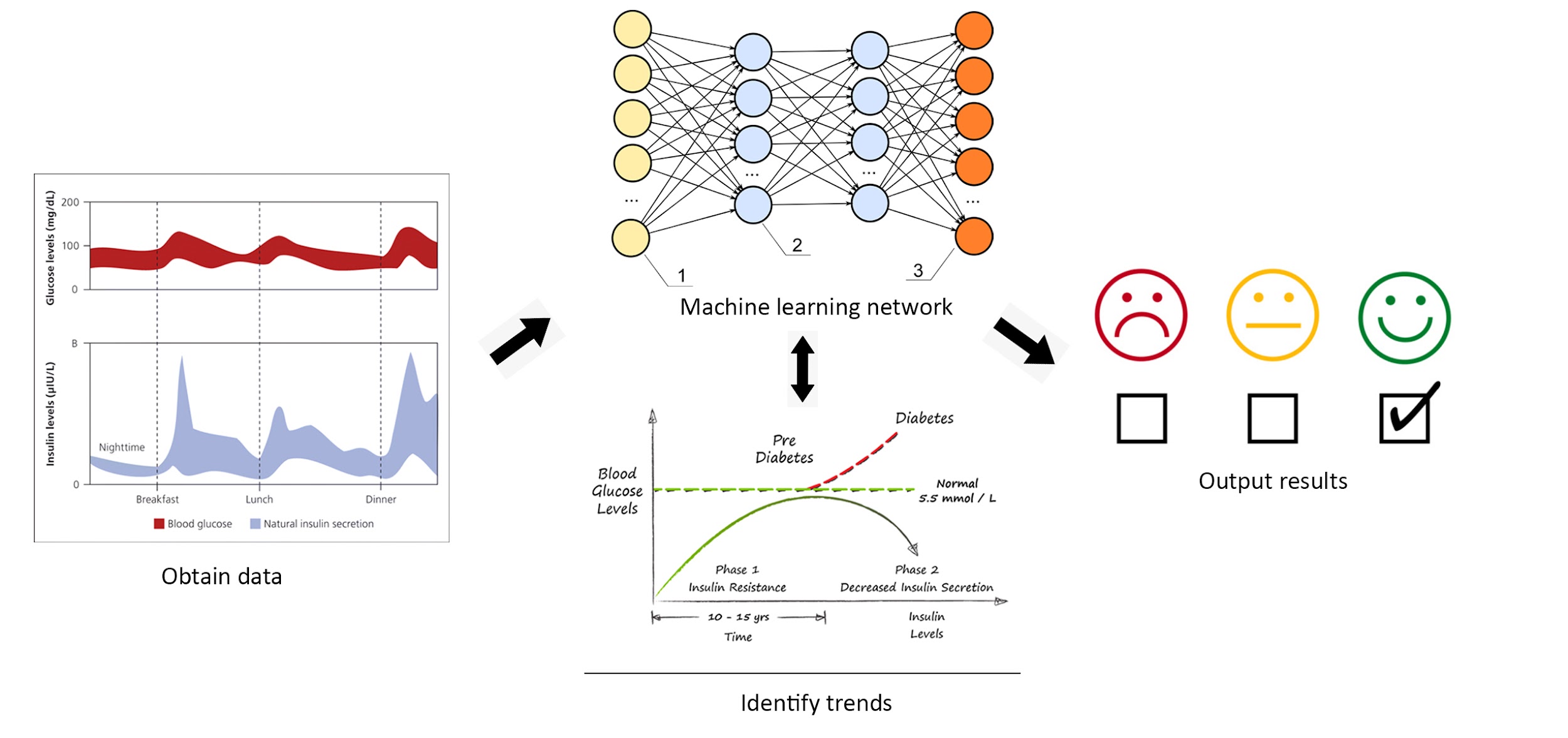
**WORD COUNT: ~1200**

### Introduction

Type 2 diabetes is an increasingly prevalent chronic health condition that arises primarily from obesity and physical inactivity [1]. This disease restricts the body’s ability to metabolize glucose [2]. Although there are currently over 2.3 million Canadians diagnosed with diabetes (around 1 in 16 people), around 20% of diabetic cases remain undiagnosed [3]. On top of living a difficult and restrictive lifestyle, individuals affected by diabetes also face a decrease in life expectancy.

Due to COVID-19, more people are transitioning to working from home, putting many at risk to an inactive and sedentary lifestyle. With hospitals full of COVID patients, diabetes may be the last on everybody’s minds. However, the risks are real, and diabetes is preventable in its early-stages. With the help of machine learning algorithms, early detection of type 2 diabetes can save many from a life-time of health complications.

### Illustration / Figure



1. Obtain data from datasets [4]
2. Import data/parameters into machine learning algorithm [5]
3. Identify trends and extrapolate information [6]
4. Output results

### Background & Related Work

Researchers from the Vellore Institute of Technology developed a machine learning based diabetes mellitus detection algorithm just last year, with a prediction accuracy of 85.6% [7]. However, this project focuses more on diabetes mellitus, a condition that leads to multi-organ failure, rather than diabetes detection as a whole. One of the datasets used was from the National Institute of Diabetes and Digestive and Kidney Diseases, which contains information on diabetic female individuals at least 21 years old of Pima Indian heritage [8]. This dataset can be found on the UCI Machine Learning Repository for public use.

There are other datasets on Kaggle, specifically Early Stage Diabetes Risk Prediction Dataset, which describes the symptoms of individuals at risk of developing type 2 diabetes [9]. Another dataset describes health-related information such as sugar levels and blood pressure of diabetic patients from Germany [10].

### Data Processing (4 points);

We will be retrieving data from Kaggle [11], the UCI Machine Learning Repository[12] and other valid dataset websites. Our data will consist of information about patients who may or may not have diabetes, some of this information may include age, body mass index (BMI), and blood pressure. There are multiple diabetes related datasets on these websites, however some of the datasets consist of the same data. To keep the dataset from containing too many duplicate entries we will research the datasets available and make sure not to add the duplicates to our dataset. For example, these datasets “Diabetes Data Set” from Kaggle and “Diabetes Data Set” from UCI Machine Learning Repository, are the exact same dataset and if we are to use their data we will make sure to only use one of them [8][13].

### Architecture

The neural network model that we will be using is the multi-layered Artificial Neural Network (ANN) model. This is because our input data is not an image but numbers about a person's health. Since there are many attributes that can determine whether someone has diabetes, our problem is nonlinear, so that is why a multi-layered ANN will be used. There will be no filtering necessary for this input data, so a Convolutional Neural Network (CNN) is not needed.

### Baseline Model

We will be using the simple linear regression model on the Pima Indians database provided by the *Towards Data Science* website as our baseline model [14]. This is a simple implementation that seeks to achieve the same goal as our project by slightly simple means. The comparison to this model will give us a good idea of whether or not our implemented neural network performed better than a simple regression model.

### Ethical Considerations

For ethical considerations we will make sure that consumer data is held with the highest level of privacy, including making sure that no data collected by this model is used by companies for any nefarious purposes (ie, insurance companies using the data to give higher rates to people who may not have diabetes, but are likely to be diagnosed later in life). We will also make sure to receive consent for all data used in this model (whether that be through using public data, etc).

### Project Plan (4 points)

Team meetings will be conducted every Thursday at 5pm, during these meetings the team will discuss the project’s progress, any challenges or issues a member is facing, and confirm the project’s future plans. Text communication through Discord and Facebook Messenger will be used as a backup. When a member has completed their assigned task, they should inform the team so team members facing issues can request assistance. The project will be organized on GitHub for easier version control.

| **Task** | **Assignee(s)** | **Responsibilities** | **Internal Deadline** |
| --- | --- | --- | --- |
| Gather training, validation and testing datasets | Sean, Priscilla | - Research datasets and collect the valid ones | February 26, 2021 |
| Implement model architecture | Sadra, Nathan | - Implement the network that was discussed during meetings | March 1, 2021 |
| Test and Train model | All | - Train the neural network  - Tune hyperparameters based on training and validation results | March. 25, 2021 |
| Progress Report | All | - Report the project’s status so far to the instructors (5%) | March 4, 2021 |
| Project Presentation | All | - Gather test results  - Complete presentation (10%) | March 28, 2021 |
| Final Deliverable | All | - Report the status of the completed project (20%) | March 31, 2021 |

### Risk Register (4 points):

| **Risk #** | **Description** | **Likelihood** | **Solution** |
| --- | --- | --- | --- |
| 1 | Determining model hyperparameters takes longer than expected | >50% | - Have team members train models with different hyperparameters.  - Seek guidance from TA/professors. |
| 2 | Team member drops the course or encounters personal problems | <10% | Distribute work evenly between available members. |
| 3 | File/Github corruption | <1% | Multiple backups of the project. |

Risk 1 describes the possibility of underestimating the amount of time it takes to tune the hyperparameters of a model. The likelihood of this risk is greater than 50% because we may not find the hyperparameters that result in a satisfactory model in a timely manner. To avoid this problem, training time will be reduced by having team members train the model with GPUs and with different hyperparameters from each other. Furthermore, TA or professor consultation will be done if model training takes much longer than expected.

Risk 2 describes the possibility of a team member becoming unable to work due to them dropping the course or encountering a personal problem. The likelihood of this happening is less than 10% because all team members are determined to complete this course. Also, the chance of personal problems keeping us from working for a long period of time is low. In case this happens the team is prepared to evenly distribute the work of the to the remaining available members.

Risk 3 describes the possibility of the main project files or github repository becoming corrupted. The likelihood of this situation is less than 1% because files and github repositories are reliable so they are not often corrupted. Each team member will also keep local copies of the project. If the situation does occur the first member to notice the corrupted files will inform the rest of the team, so that the corrupted files are not shared.

### Link to GitHub

GitHub Link: <https://github.com/NathanJones98/APS360_Team_62>

### References

[1] “Physical inactivity and diabetes,” World Health Organization, 12-Nov-2015. [Online]. Available: <https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/news/news/2015/11/physical-inactivity-and-diabetes>. [Accessed: 11-Feb-2021].

[2] “What is diabetes?,” DiabetesCanadaWebsite. [Online]. Available: <https://www.diabetes.ca/about-diabetes/what-is-diabetes>. [Accessed: 11-Feb-2021].

[3] P. H. A. of Canada, “Government of Canada,” Highlights: in Canada: Facts and figures from a public health perspective - Canada.ca, 15-Dec-2011. [Online]. Available: <https://www.canada.ca/en/public-health/services/chronic-diseases/reports-publications/diabetes/diabetes-canada-facts-figures-a-public-health-perspective/report-highlights.html>. [Accessed: 11-Feb-2021].

[4] G. B. M. Gibbs and M. Gibbs, “Ersatz offers Machine Learning as a Service,” Network World, 23-Jan-2014. [Online]. Available: <https://www.networkworld.com/article/2226201/ersatz-offers-machine-learning-as-a-service.html>. [Accessed: 11-Feb-2021].

[5] “Understanding Insulin Resistance is the First Step to Reversing Diabetes:” WELLNESS GARAGE. [Online]. Available: https://www.wellnessgarage.ca/better-blog/understanding-insulin-resistance-is-the-first-step-to-reversing-diabetes. [Accessed: 11-Feb-2021].

[6] “Glossary,” Diabetes Education Online. [Online]. Available: <https://dtc.ucsf.edu/types-of-diabetes/type2/treatment-of-type-2-diabetes/medications-and-therapies/type-2-insulin-rx/insulin-basics/>. [Accessed: 11-Feb-2021].

[7] J. Chaki, S. T. Ganesh, S. K. Cidham, and S. A. Theertan, “Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review,” Journal of King Saud University - Computer and Information Sciences, 04-Jul-2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1319157820304134>. [Accessed: 11-Feb-2021].

[8] UCI Machine Learning Repository: Diabetes Data Set. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/diabetes>. [Accessed: 11-Feb-2021].

[9] I. Dutta, “Early Stage Diabetes Risk Prediction Dataset,” Kaggle, 21-Sep-2020. [Online]. Available: <https://www.kaggle.com/ishandutta/early-stage-diabetes-risk-prediction-dataset>. [Accessed: 11-Feb-2021].

[10] John, “diabetes,” Kaggle, 25-Apr-2018. [Online]. Available: <https://www.kaggle.com/johndasilva/diabetes>. [Accessed: 11-Feb-2021].

[11] “Find Open Datasets and Machine Learning Projects,” Kaggle. [Online]. Available: <https://www.kaggle.com/datasets>. [Accessed: 11-Feb-2021].

[12] UCI Machine Learning Repository: Data Sets. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets.php>. [Accessed: 11-Feb-2021].

[13] M. A., “Diabetes Data Set,” Kaggle, 05-Aug-2020. [Online]. Available: <https://www.kaggle.com/mathchi/diabetes-data-set>. [Accessed: 11-Feb-2021].

[14] K. Dhandhania, “End-to-End Data Science Example: Predicting Diabetes with Logistic Regression,” *Medium*, 24-May-2018. [Online]. Available: https://towardsdatascience.com/end-to-end-data-science-example-predicting-diabetes-with-logistic-regression-db9bc88b4d16. [Accessed: 11-Feb-2021].