FROM DATA TO DIABETES PREDICTION: UNMASKING PATTERNS WITH MULTILAYER PERCEPTRON NETWORKS USING NUTRITIONAL AND WELLNESS FACTORS

Kristjan Meitz, Ureem James and Jagadish Velaga

August 2nd, 2023

Addresses to which comments should be addressed
Kristjan Meitz: kcmeitz@gmail.com
Ureem James: ureemjames18@gmail.com
Jagadish Velaga: jvelaga1@gmail.com

Abstract

Given the push in early screening for diabetes, it is vital that medical professionals are equipped with the most accurate tools to predict the likelihood of their patients having diabetes. Support Vector Machines and Tree based methods have been commonly used to predict diabetes amongst patients with strong accuracy. In contrast, Neural Networks, though explored, have been underutilized in early screening methods for predicting diabetes. The purpose of this paper is to develop a Multilayer Perceptron (MLP) model to predict diabetes and compare the performance of the model against commonly applied approaches. MLPs are effective in eliminating less relevant data and features and identifying complex linear and nonlinear relationships. The dataset utilized contains many complex demographic, dietary, medical, and socioeconomic factors, making a MLP model an attractive method for diabetes prediction. The MLP model was compared against three other machine learning methods (Support Vector Machine, Random Forest, and XGboost) and performed the best with an accuracy of 91% and recall of 71%.

Keywords: Diabetes Prediction, Healthcare Analytics, Multilayer Perceptron, Support Vector Machine, Random Forest, XGboost, Machine Learning, Artificial Intelligence, Kernel Methods

Introduction

The purpose of this paper aims to explore the performance of a Multilayer Perceptron Neural Network (MLP) on a large structured diabetes dataset with numerous features, and to contrast its performance against three commonly applied and successful machine learning models in diabetes prediction: Random Forests (RF), XGBoost (XGB) and Support Vector Machines (SVM).

Timely detection of diabetes is vital in initiating prompt medical care and saving lives. In March 2021, NBC News reported that the Preventive Services Task Force recommends early screening for diabetes, starting as early as age 35, five years ahead of the previously recommended age of 40, to ensure more people undergo screening at an earlier age to avoid the potentially fatal implications of undetected diabetes (Carroll 2021).

Consequently, the heightened public awareness for early diabetes screening has motivated researchers to predict the likelihood of diabetes with greater accuracy than ever before, utilizing both classical prediction and deep learning models. However, most researchers encounter a common challenge when working with healthcare data: large datasets with potentially irrelevant information and complex nonlinear relationships. As a response to these limitations of classical machine learning models, researchers have turned to neural networks to predict the likelihood of diabetes. Neural networks have the capacity to model complex and hidden relationships, both linear and nonlinear, within large structured and unstructured datasets with numerous features. Though neural networks have existed since 1958, their usage has become more popular since the turn of the 21st century due to the ubiquity of big data and greater computing power; their impressive performance specifically in speech and image recognition tasks; and their adoption as the industry standard in products offered by Google, Microsoft, Apple, etc. (Macukow 2016).

Therefore, this paper will explore the performance of a MLP on a large dataset sourced from the National Health and Nutrition Examination Survey (NHANES) to predict the likelihood of diabetes.

Literature Review

In recent years, the literature on the application of machine learning and deep learning in diabetes has witnessed significant growth, revolutionizing medical research and diagnosis as a whole. This research ranges from image analysis and processing, analysis of blood sugar levels, wearable devices, to genomics. However, this literature review will specifically focus on the use of machine learning and neural networks to predict whether an individual has diabetes utilizing dietary, medical, and socioeconomic factors.

Figure 1

Machine learning and MLP literature review summary

Author(s) (Year)	Dataset	Algorithms: Accuracy
Dinh et al. (2019)	NHANES	● XGB: 86.2% ● RF: 85.6% ● SVM: 84.9% ● Logistic Regression (LR): 82.7%
Qin et al. (2022)	NHANES	● CATBoost: 82.1% ● RF: 78.4% ● XGB: 70.8% ● LR: 68.9% ● SVM: 67%
Giveki et al. (2012)	UIC	● IM-MCS-FWSVM: 92.6% ● PCA-PSO-LS-SVM: 82.8% ● PCA-MI-LS-SVM: 80.0% ● PCA-LS-SVM: 79.2%
Chari et al. (2019)	Central Virginia	● RF with Feature Selection: 92% ● RF: 85.6% ● Bagging with Decision Tree: 81.3% ● Decision Tree: 75.2%
Sivasankari et al. (2022)	PIMA	● MLP: 86.1%

Author(s) (Year)	Dataset	Algorithms: Accuracy
Abushawish and Nassif (2023)	Bangladesh Hospital	● KNN: 98.0% ● Radial-Basis Neural Network: 96.0% ● SVM: 94.0% ● MLP: 91.0%
Mishra et al. (2020)	PIMA	● EA-GA-MLP: 98% ● A-GA-MLP: 93.2% ● E-GA-MLPL: 94.1% ● GA-MLP: 92.3%
H. Temertas et al. (2009)	PIMA	● MLP: 79.6% ● SVM: 79.2% ● LS-SVM: 78.2% ● PNN: 78.1%
Butt et al. (2020)	PIMA	● MLP: 86.1% ● RF: 77.4% ● LR: 73.1%
Salameh et al. (2021)	Unknown	● MLP: 77.6% ● SVM/KNN had lower accuracy, scores not provided

Machine Learning Approaches

In the article, *A data-driven approach to predicting diabetes and cardiovascular disease with machine learning,* the researchers created machine learning models to predict if patients have diabetes and cardiovascular disease, along with identifying key factors that contribute to developing these diseases (Dinh et al. 2019). Researchers used the NHANES dataset to build LR, SVMs, RF, and gradient boosting models. For each model, they applied a weighted ensemble approach through a 10-fold cross-validation. They performed feature selection to reduce the dimensionality of the data and relied on the XGB ensemble classifier to choose the top 24 features. The top non-lab features were waist, age, self-reported greatest weight, and leg length. The top lab diabetes classifiers with lab results were blood osmolality, sodium, and blood urea nitrogen. The results showed that the gradient boosting models outperformed all the other models based on the AU-ROC, precision, recall, and F1 scores. The LR model had the worst performance of all the models. One potential limitation of this research is that it only relied on the XGB ensemble classifier for feature selection. This resulted in the XGB model achieving the best results.

The National Institutes of Health (NIH) conducted similar research in a study titled *Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type* (Qin et al. 2022). Researchers in this study utilized the NHANES dataset to predict diabetes by incorporating lifestyle factors and demographic data. The researchers used five machine learning methods:

- CATBoost
- XGB
- RF
- LR
- SVM

Considering the class imbalance in the dataset, the study employed SMOET_NC (Synthetic Minority Over-sampling Technique Nominal Continuous) to address the imbalanced data. For feature selection, the researchers employed 18 variables commonly used for predicting diabetes and introduced additional variables to enhance variability. They employed stepwise selection to finalize the variables and evaluated each model based on the Precision, Sensitivity, Specificity, and F1 score. The CATBoost model achieved the highest accuracy score of 82.1% and an AUC of 0.83.

In the research paper, *Automatic Detection of Diabetes Diagnosis using Feature Weighted Support Vector Machines based on Mutual Information and Modified Cuckoo Search*, the researchers constructed a Feature Weighted SVM model to predict diabetes in the UIC dataset, which included 768 observations and 8 medical and demographic features (Giveki et al. 2012). Kindly refer to the Appendix A for more information on the dataset. To discover the most relevant and informative subset of features, the researchers utilized Principal Component Analysis (PCA) to reduce the dimensionality of the data. To consider the varying contributions of different features to the classification of diabetes, the researchers employed Mutual Information to assign weights to the features in the model. The SVM model achieved the highest performance, with an overall accuracy score of 93.6%, following the application of a Modified Cuckoo Search (MCS) to expedite convergence.

In the research paper, *Classification of Diabetes using Random Forest with Feature Selection Algorithm*, the researchers constructed various machine learning models to predict the presence of diabetes using demographic and medical data (Chari, Babu, and Kodati 2019). The study analyzed a sample dataset that included 403 observations and 19 features from African Americans in central Virginia. The goal was to evaluate the prevalence of obesity, diabetes, and other cardiovascular risk factors. The study evaluated the performance of the following machine learning algorithms:

- Decision Tree
- Bagging with Decision Tree
- RF with and without Feature Selection

Among all the models selected for comparison, the RF with feature selection achieved the highest accuracy score of 92%.

MLP Approaches

In their study, Sivasankari et al. (2022) utilized the PIMA Indian Diabetes dataset, which contains 768 observations of female patients over the age of 21 and 8 features. They developed a MLP with eight input neurons corresponding to each variable in the dataset, a single hidden layer, and a sigmoid activation function to classify diabetes diagnoses. The researchers then compared the results of the MLP model to other traditional machine learning methods and the MLP had the best performance across each metric with an accuracy score of 86.08%.

Similarly, in 2023, Abushawish and Nassif conducted a study involving 520 hospital patients in Bangladesh to predict their diabetic status. The researchers utilized 15 features focused on demographic information and health symptoms. They compared the performance of the following models:

- MLP
- Radial-Basis Neural Network (RBF)
- SVM
- K-Nearest Neighbors (KNN)

Each model exhibited remarkably high accuracy scores and recall rates, but the KNN model outperformed the others, while the MLP model demonstrated the weakest performance. A more robust and larger dataset with intricate relationships between input and output variables would likely enhance the performance of an MLP model, making it better suited for the specific characteristics of the dataset we are employing.

In the study, *EAGA-MLP—An Enhanced and Adaptive Hybrid Classification Model for Diabetes Diagnosis*, researchers employed the PIMA Indian Diabetes dataset to develop an MLP model for predicting diabetes (Mishra et al. 2020). To overcome the limitations of the small dataset, they devised an attribute optimization algorithm, Enhanced and Adaptive Genetic Algorithm (EAGA), to generate new samples. The researchers highlighted the rising complexity of diagnosing and treating diabetes due to increased variation in diabetic cases and intricate medical patient data. While machine learning and data mining techniques offer promise for early diabetes detection, the abundance of irrelevant and ambiguous factors in these datasets has hindered the effectiveness of traditional machine learning classification algorithms.The researchers contend that the EAGA-MLP model can tackle the dataset complexity by identifying relevant attributes and important symptoms, thereby reducing data size and complexity without omitting crucial factors. Study outcomes underscored the EAGA-MLP model's superiority over each employed classification model, achieving an impressive accuracy score of 97.76%.

In a study titled *A comparative study on diabetes disease diagnosis using neural networks,* the researchers explored the PIMA diabetes diagnosis dataset employing both a probabilistic neural network and an MLP architecture (H. Temertas, Yumusak, and F. Temurtas 2009). The MLP configuration comprised 50 neurons for each hidden layer and an output layer using a sigmoid activation function. Conversely, the probabilistic neural network featured a single radial basis hidden layer with locally tuned units, linked to a two-unit output layer. This hidden layer employed radial basis functions to compute the Euclidean distance from the center to the input vector, subsequently applying the radial basis function. Outcome comparison revealed a slightly superior accuracy of the MLP model over the probabilistic neural network. Benchmarking against related studies demonstrated average model performance. However, the researchers concluded that rapid convergence of the models led to overfitting due to a memorization effect.

In the research paper *Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications*, the researchers developed predictive models for diabetes detection within the PIMA dataset (Butt et al. 2020). They employed the following machine learning and neural network methods:

- RF
- MLP
- LR

The researchers leveraged an IoT-based monitoring system and monitored blood glucose levels to predict diabetes. The IoT-based monitoring system used the following models for diabetic forecasting:

- Long Short-Term Memory Neural Network
- Moving Averages
- Linear Regression

Evaluation metrics included recall, precision, and accuracy for classification models, along with correlation, root mean square error, and accuracy for diabetic forecasting models. Notably, the MLP model excelled among classification models, while the LSTM model demonstrated superior performance in diabetic forecasting.

In the study *Prediction of Diabetes and Hypertension using Multi-layer Perceptron Neural Networks*, the researchers evaluated the efficacy of artificial neural networks (ANNs), particularly MLP, in forecasting diabetes and blood pressure disorders (Salameh et al. 2021). Medical and demographic factors were integrated into each model for diabetes prediction. The

MLP model demonstrated superior performance over other classifiers, including SVM and KNN, achieving an accuracy score of 77.6% for diabetes and 68.7% for hypertension.

In conclusion, this literature review highlights the use and effectiveness of machine learning methods and MLP neural networks in predicting diabetes using dietary, demographic lifestyle, and medical factors. It is clear that these methods and architectures give valuable insights and applications for improving healthcare and patient outcomes. The SVM, RF, and XGB models outperformed other machine learning methods, but MLPs were also effective in diagnosing diabetes and were often more accurate in predicting diabetes than machine learning classifiers. The MLP models exhibit enhanced performance when dealing with datasets that are larger and more complex.

Data

We utilized the NHANES data, a pre-pandemic survey conducted by the National Center for Health Statistics, and divided the data into four sections:

- Demographics Data
- Dietary Data
- Examination Data
- Laboratory Data

Each survey participant was designated a unique identifier - Respondent Sequence Number (SEQN), which served as the primary key to join all the relational data files. After the join, we observed that:

- An average of 34.72% of all observations were missing across all features.
- A class imbalance existed in our supervisor variable - only 9.6% had diabetes.

The disproportionate representation of labels in our main supervisor variable raised concerns about developing proportionate training and test datasets. That said, upon further analysis of the missing values, we found that:

- 39% of all missing observations were concentrated between ages of 1 and 18
- Only 3% of all observations between the ages of 1 and 18 had diabetes, whereas 40% did not have diabetes.
- 37 features had at least 60% missing observations.

Given the constraints to impute for missing values, we removed the 37 features from the dataset, and removed all observations whose respondents were below the age of 18. The decision not only greatly reduced the percentage of missing values, but also rebalanced our supervisor variable, resulting in 20% of all observations being labeled as having diabetes. We then proceeded to cleanse the remaining data of outliers and imputed missing values using the median for numerical features and the mode for categorical features, yielding a final dataset comprising 9459 observations and 133 features.

Methods

We selected RF, SVM, XGB and MLP as our choice of classification algorithms. Our choice was derived from those we found commonly employed in solving similar problems in our Literature Review section. The table below shows the benefits and important characteristics of each model in predicting diabetes.

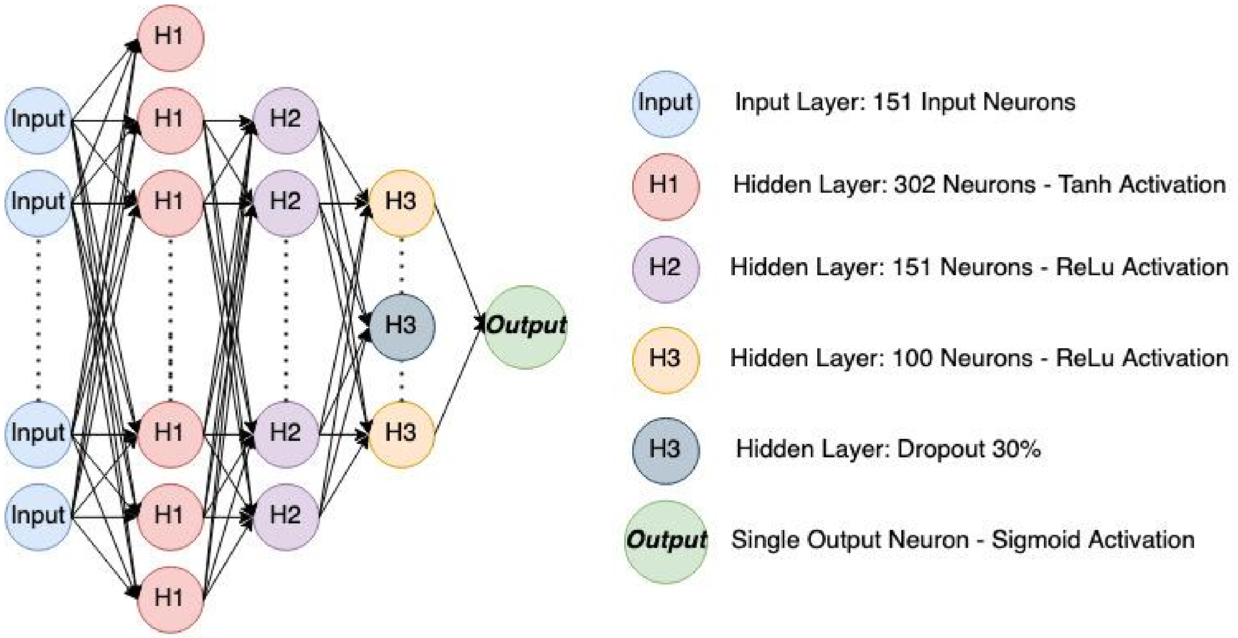
Figure 2

Comparative analysis of algorithm characteristics, literature, and accuracy

Algorithm	Important Characteristics	Literature Using the Algorithm and Accuracy Score
MLP	● Identifies linear and non-linear relationships ● Effective feature learning from disparate datasets ● Drop out and weight decay can help prevent overfitting ● Handles missing data well	● Sivasankari et al. 2022: 86.1% ● Abushawish and Nassif 2023: 91% ● Mishra et al. 2020: 94.7% ● H. Temertas et al. 2009: 82.4% ● Butt et al. 2020: 86.1% ● Salameh et al. 2021: 77.6%
SVM	● Effective with small sample sizes ● Identifies linear and non-linear relationships ● Handles high-dimensional data well ● Effective feature selection	● Dinh et al. 2019: 84.9% ● Qin et al. 2022: 83.9% ● Giveki et al. 2012: 93.6% ● Abushawish and Nassif 2023: 94%
RF	● Ensemble of decision trees reduces overfitting ● Identifies linear and non-linear relationships ● Robustness to outliers ● Provides feature importance scores	● Dinh et al. 2019: 85.5% ● Qin et al. 2022: 84.4% ● Chari et al. 2019: 92% ● Butt et al. 2020: 77.4%
XGB	● Avoids overfitting through regularization ● Automatically captures complex relationships and interactions between features ● Handles class imbalance effectively ● Provides feature importance scores	● Dinh et al. 2019: 86.2% ● Qin et al. 2022: 83%

Figure 3

Selected Multilayer Perceptron Model


Input
H1
H2
H3
Output
Input Layer: 151 Input Neurons
Hidden Layer: 302 Neurons - Tanh Activation
Hidden Layer: 151 Neurons - ReLu Activation
Hidden Layer: 100 Neurons - ReLu Activation
Hidden Layer: Dropout 30%
Single Output Neuron - Sigmoid Activation

***Note**: The ellipses between the neurons in the same layer represent additional neurons*

Figure 3 illustrates our selected MLP model architecture. We arrived at 151 input neurons after one-hot encoding our categorical variables and scaling our feature variables. Our model was trained using Stochastic Gradient Descent (SGD) with a learning rate of 0.001 and a batch size of 128 samples across 200 epochs.

Selected Machine Learning Models

SVM, RF and XGB were all tuned for the best parameter settings using 5-folds cross validation. The following notable settings were used for the models:

1) SVM used a non-linear kernel transformation - Radial Basis Function - alluding to a non-linear relationship between our predictors and class labels.
2) RF was tuned by using hyperparameters - max-features at each node, max depth, max features, and minimum samples at the leaf.

3) XGB underwent tuning by adjusting the learning rates, number of tree estimators, and the maximum depth of each tree ensemble. To enhance feature selection, we employed the SelectFromModel methods available in Scikit-learn.

Selected Scoring Metrics

We employed commonly used classification scoring metrics to assess the performance of our models:

- Accuracy
- Precision
- Recall
- Specificity
- F1 Score

Additionally, between Precision and Recall, *more preference will be given to Recall* given that it's vital we accurately capture the proportion of samples that are actually diabetic.

Results

Figure 4 below shows performance metrics of the selected models run on the diabetes dataset. MLP and ML model results are presented in tabular form for comparison. Details will be covered in the Analysis and Interpretation section.

Figure 4

MLP and ML model performance metrics

Model	Accuracy	Precision	Recall	Specificity	F1-Score
MLP	91.0%	75.0%	71.0%	94.5%	72.9%
RF	90.6%	83.6%	56.6%	98.0%	68.0%
SVM	91.0%	83.8%	56.0%	98.0%	67.0%
XGB	90.6%	78.9%	62.1%	96.5%	69.5%

Analysis and Interpretation

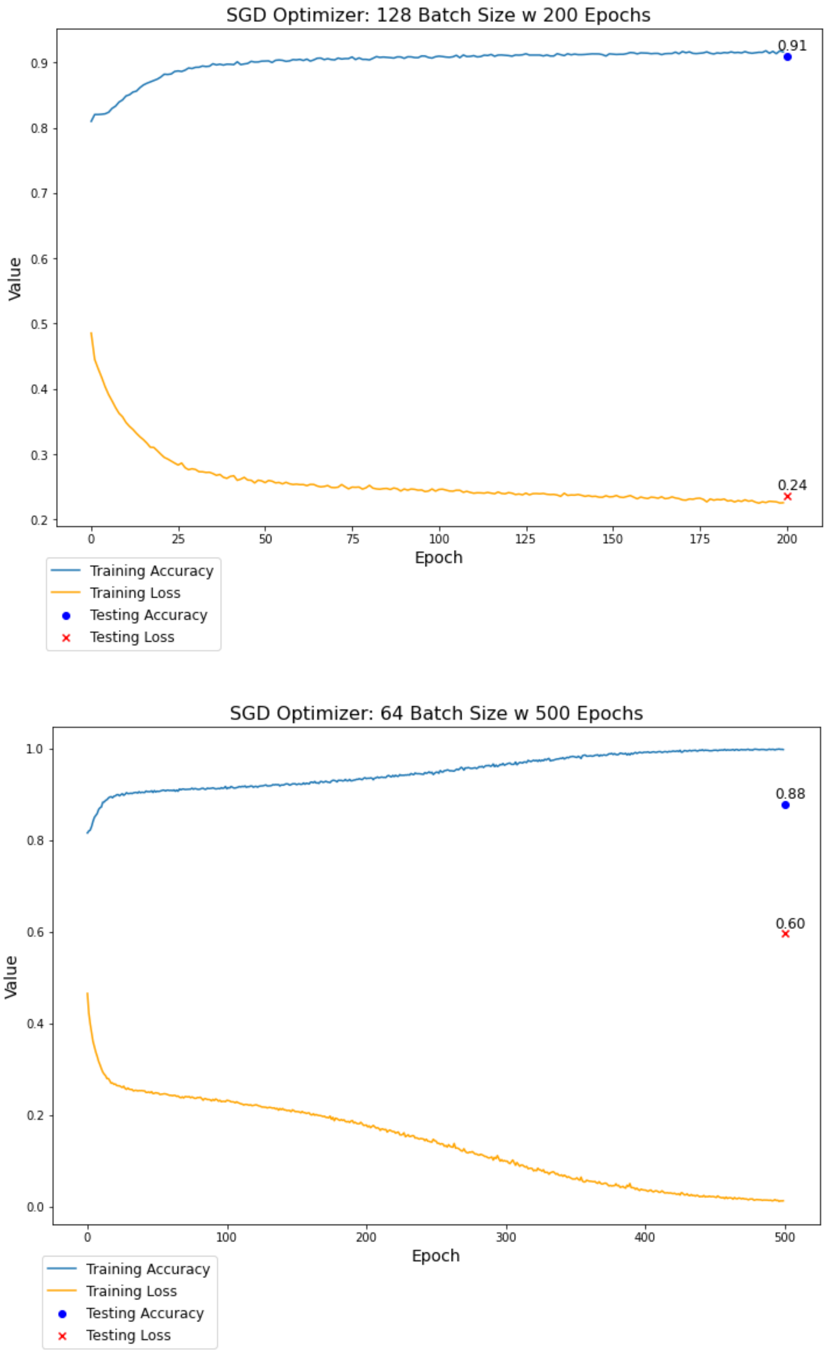
Model Performance and Settings

The models performed similarly in terms of accuracy. That said, the MLP model had the best overall results - performing significantly better in terms of Recall, which outlines the percentage of Diabetic samples that were correctly predicted out of all Diabetic samples. However, achieving performance metrics wasn’t solely a product of our model choice. In our pursuit to optimize the MLP model’s performance, we tested various optimization strategies discussed below.

Halting training at 200 epochs with a batch size of 128 samples, led to better generalization even though additional epochs were tested with a smaller batch size of 64 samples. This could be due to the model overfitting as it progressed beyond 200 epochs, which would hinder generalization. Kindly refer to the visuals below that showcase worse generalization with more iterations and fewer batch sizes.

Figure 5

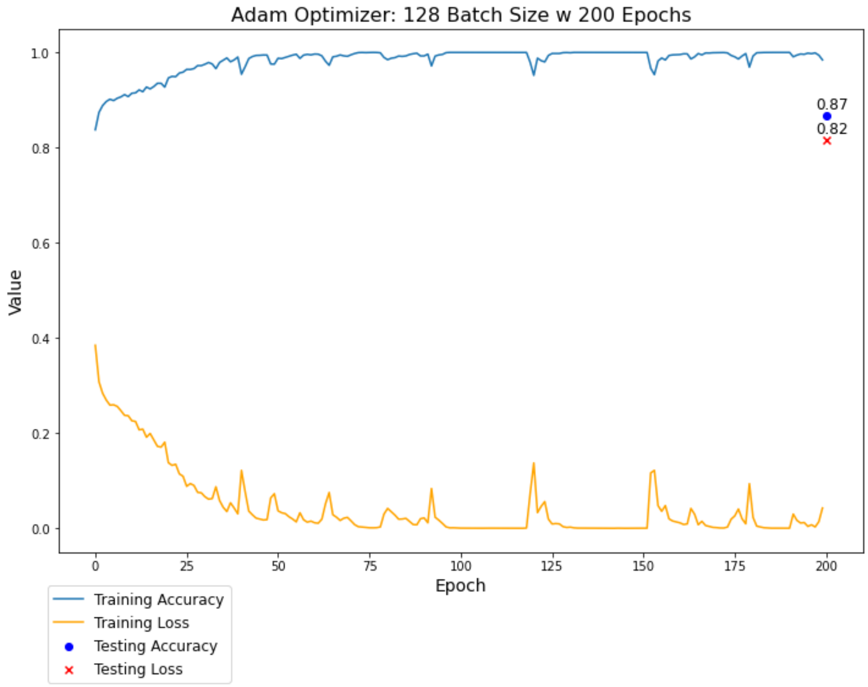
Assessing the impact of more epochs and smaller batch sizes



A few optimization strategies were tested, out of which SGD generalized the best. In contrast, an Adam optimization model outperformed the SGD model on the training data, but didn't generalize as well as the SGD model (refer to visual below).

Figure 6

Assessing the Adam Model Performance



There could be a few contributing factors:

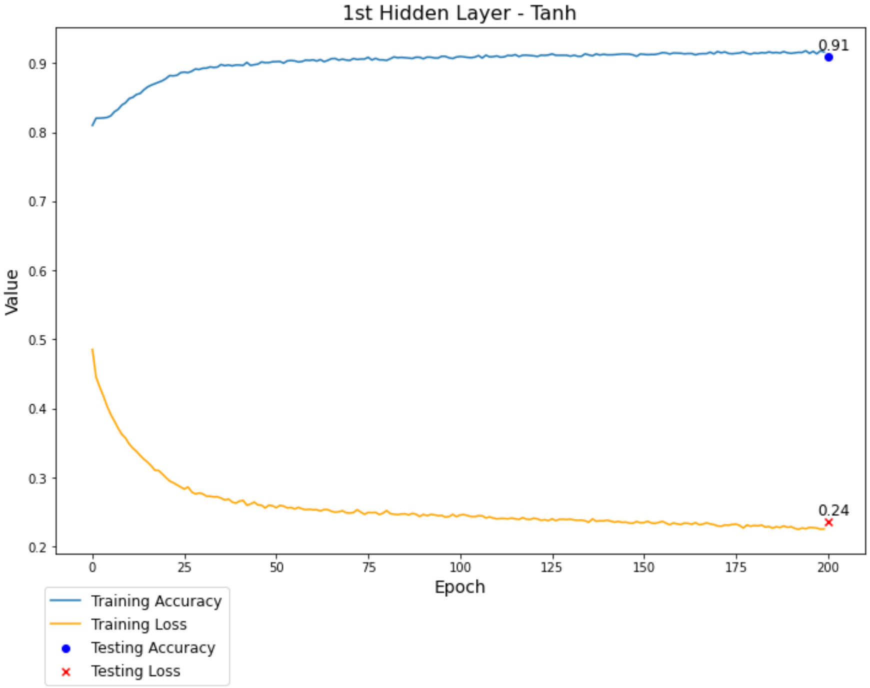
- The frequent peaks in the loss curve in the chart indicate that the model has been overshooting the optimal solution frequently. This would imply that the learning rate, currently the same as in SGD, was too high and needs adjustment.
- The randomness of SGD, which is an inherent regularizer, can help better avoid overfitting and therefore lead to better generalization (Lei et al. 2018).

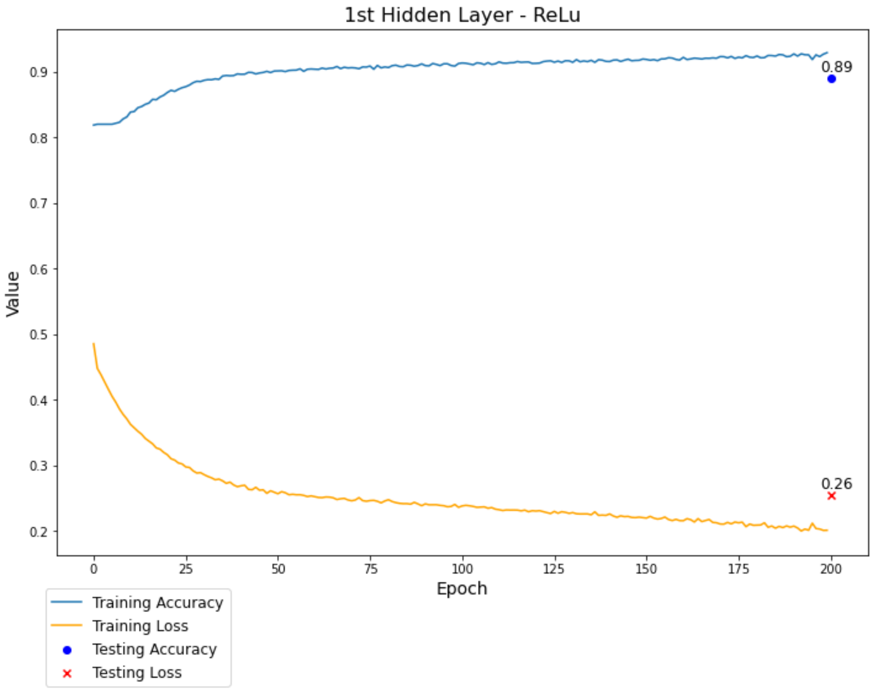
- The peaks in the loss curve could also suggest that the Adam model was inefficient in capturing the high variability in the data since smaller batch sizes provide less information, requiring larger weight updates after each epoch.

Including a Hyperbolic Tangent activation function instead of a ReLu activation function as the first hidden layer, increased generalization. Kindly refer to the visuals below.

Figure 7

Assessing the impact of the a ReLu Activation in the first Hidden Layer





This could be due to a few reasons - the smoothness of the hyperbolic tangent function leading to more consistent convergence, or perhaps since some neurons were consistently outputting close to zero values, this could've introduced "the dying neuron" effect whilst using ReLu and hampered loss learning as can be seen from the loss curve being less steep in epochs 0 to 25 in the second visual.

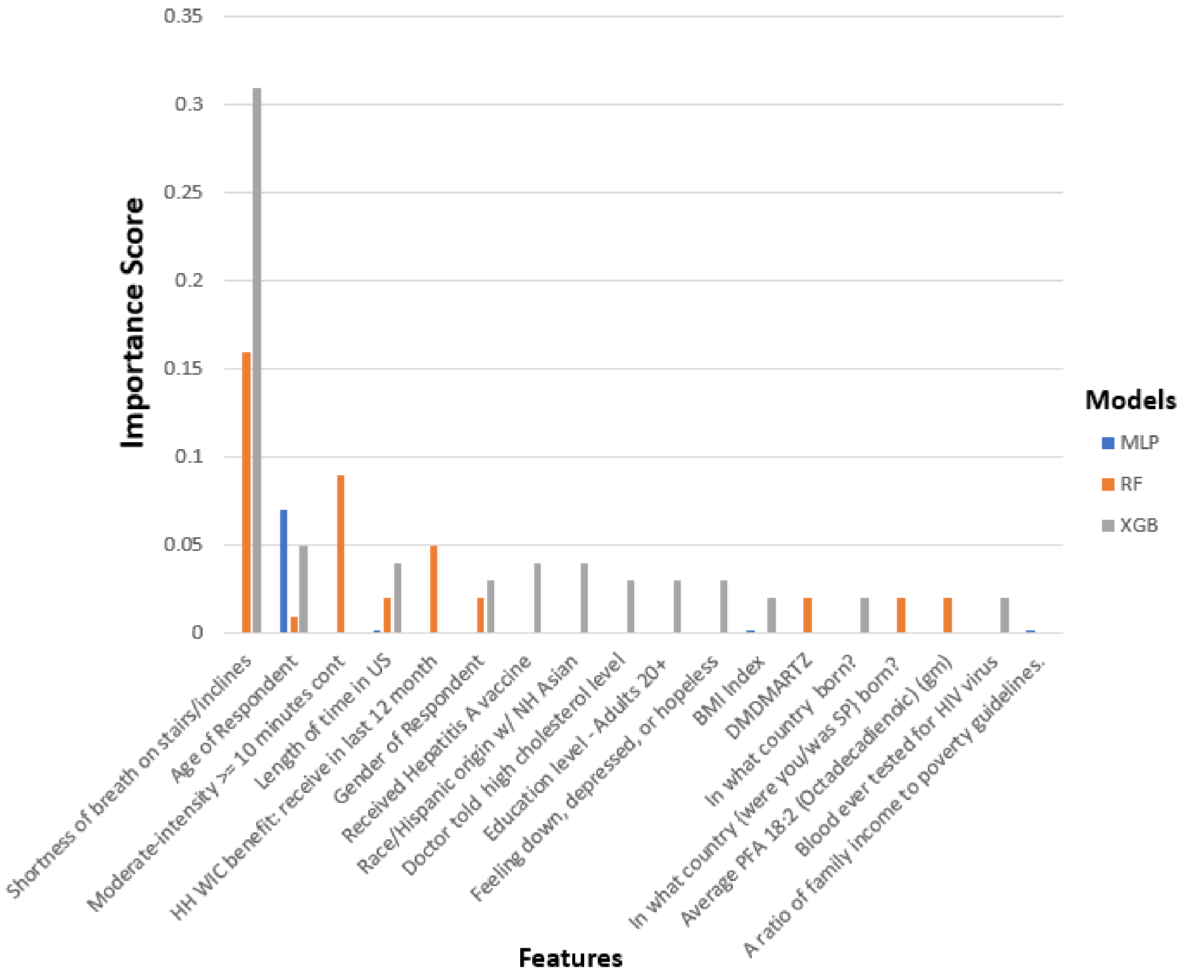
Finally, we tested including a drop out parameter to the third hidden layer, which switched off 30% of neurons at random. This provided consistent accuracy scores across all seeds that were tested. Kindly refer to the Directions for Future Work section where we outline potential improvements and suggestions.

Feature Importance

Identifying factors that propagate diabetes is key in predicting diabetes. Figure 5 shows a consolidated view of the most important features and scores from our MLP, XGB and RF models. We weren’t able to retrieve the important features from the SVM model because we used a non-linear kernel transformation to transform the features during training.

Figure 5

Features importance by model



In comparing the important features of each model, we observed some similarities and differences. We attribute the presence of differences to the MLP model being able to extract deeper non-linear patterns whereas the ML models are a form of shallow learning. For example, Age was identified as the most important feature by the MLP model, whereas ‘Shortness of breath on stairs/incline’ was preferred by XGB and RF. Though different, this is not entirely alarming as we can intuitively expect the two to be correlated. Furthermore, all important features identified by our MLP model were also identified by the XGB and RF models. MLP had very low importance scores for all features, barring ‘Age’. All the features had an importance score of less than 1/10 of the top feature score. This could be either because most features are noisy or the impact of some features is muted by the “dying neuron” effect coming from our ReLu layers. In contrast, XBG and RF identified multiple similar features with similarly high scores, which could be due to the fact they are finding local patterns, whereas the MLP is finding global patterns that could potentially reside much deeper within the data.

We are also surprised that the models did not pick laboratory features like Insulin, Blood Osmolarity, Glucose, etc., which were strong predictors that we came across in our literature review. It could be that wellness and socioeconomic factors are more influential in spreading diabetes, whereas the laboratory results were likely muted by individuals successfully managing diabetes.

To assess reasonability, we compared the important features identified by our models against commonly known reasons for diabetics - age, obesity, poverty, etc. Our results indicate that the models performed well by identifying similar reasons. All models identified ‘Age’ as a key contributor, even though XGB and RF identified a confounding variable. This wasn’t surprising given that our data showed 80% of people 55 or older had diabetes.

Similarly, aligning with obesity, our models identified ‘BMI Index’ as another important feature. Our models also identified the ‘'length of the time in the US’ as an important feature, which isn’t intuitive from the onset. But, given that US residents rank high in obesity statistics, the ‘length of time in the US’ could be having a confounding effect. Finally, all models identified features that aligned with poverty. The MLP model identified the ‘Ratio of family income to poverty guidelines’, whereas the RF model identified ‘HH WIC benefits received in the last 12 months’.

Conclusions

Early detection of diabetes is essential in minimizing health risks and improving patient outcomes. This work demonstrates the application of MLP models for classification and prediction for diabetes. The proposed MLP model is compared against traditional machine learning techniques including RF, SVM, and XGB which have been commonly used in similar literature. The performance of the MLP models was very similar to the machine learning models, however, the MLP model had a better recall score, which is essential for screening and early detection of the disease. Furthermore, our model's performance, as demonstrated through comprehensive performance metrics, has shown its potential as a formidable tool in aiding medical practitioners and public health professionals in early diabetes detection and risk assessment.

Directions for Future Work

- Continued experimentation with MLP architectures, including more experiments with different numbers of hidden layers and neurons along other activation functions.
- Experiment with different methods of handling the class imbalance in the data, such as different loss functions and over and undersampling methods, potentially improving our Recall score.
- Expand the dataset to include more years of data beyond the 2017 to 2020 timeframe used in this analysis to work with a more complete and rich dataset.

- Improve generalization and robustness by evaluating the model on other datasets to remove potential biases. It would be beneficial to test the models on different populations and settings.
- Adjusting the learning rate and learning rate schedule, which decreases learning over time, could lead to improved results using both the Adam and SGD model.
- Given the high variability in our data, we would also want to test our Adam model with higher batch sizes given that smaller batch sizes provide the model less information, leading to larger updates in weights after each iteration.
- Along with other regularization techniques, adding additional drop out parameters within the hidden layers within the Adam model to see if generalization can be improved.

References

Abirami, S, and P Chitra. “Energy-Efficient Edge Based Real-Time Healthcare Support System.” Advances in Computers, October 23, 2019. https://www.sciencedirect.com/science/article/pii/S0065245819300506.

Abushawish, Abdulaziz Y.I, and Ali Bou Nassif. “Prediction Of Early-Stage Diabetes Using Machine Learning.” IEEE Xplore, February 2023. https://ieeexplore.ieee.org/abstract/document/10180804.

Arya, Monika, Hanumat Sastry G, Anand Motwani, Sunil Kumar, and Atef Zaguia. “A Novel Extra Tree Ensemble Optimized DL Framework (ETEODL) for Early Detection of Diabetes.” Frontiers in public health, February 15, 2022. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8885585/.

Bani-Salameh, Hani, Shadi M. Alkhatib, Moawyiah Abdalla, Mo’taz Al-Hami, Ruaa Banat, Hala Zyod, and Ahed J. Alkhatib. “Prediction of Diabetes and Hypertension Using Multi-Layer Perceptron Neural Networks.” World Scientific. January 4, 2021. https://www.worldscientific.com/doi/abs/10.1142/S1793962321500124?journalCode=ijmssc.

Butt, Umair Muneer, Sukumar Letchmunan, Mubashir Ali, Fadratul Hafinaz Hassan, Anees Baqir, and Hafiz Husnain Raza Sherazi. “Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications.” Journal of healthcare engineering, September 29, 2021. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8500744

Carroll, Linda. “New Diabetes Guidelines Lower Screening Age to 35 for Some Adults.” NBCNews.com, March 16, 2021. https://www.nbcnews.com/health/diabetes/new-diabetes-guidelines-lower-screening-age-35-some-adults-n1261225.

Centers for Disease Control and Prevention “2017-March 2020 Pre-Pandemic Demographics Data - Continuous Nhanes.” Accessed August 2, 2023. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&Cycle=2017-2020.

Chari, K.Koteswara, M.Chinna babu, and Sarangarm Kodati. “Classification of Diabetes Using Random Forest with Feature Selection Algorithm.” IJITEE, November 2019. https://www.ijitee.org/wp-content/uploads/papers/v9i1/L35951081219.pdf.

Dinh, An, Stacey Miertschin, Amber Young, and Somya D. Mohanty. “A Data-Driven Approach to Predicting Diabetes and Cardiovascular Disease with Machine Learning - BMC Medical Informatics and Decision Making.” BioMed Central, November 6, 2019. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0918-5.

E P, Prakash, Srihari K, S Karthik, Kamal M V, Dileep P, Bharath Reddy S, Mukunthan M A, et al. “Implementation of Artificial Neural Network to Predict Diabetes with High-Quality Health System.” Computational intelligence and neuroscience, May 30, 2022. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9170457/.

Giveki, Davar, Hamid Salimi, GholamReza Bahmanyar, and Younes Khademian. “Automatic Detection of Diabetes Diagnosis Using Feature Weighted Support Vector Machines Based on Mutual Information and Modified Cuckoo Search.” Arxiv, January 2012. https://arxiv.org/ftp/arxiv/papers/1201/1201.2173.pdf.

IBM. “What Is Random Forest?” Accessed August 20, 2023. https://www.ibm.com/topics/random-forest

Koehrsen, Will. “Random Forest in Python.” Medium, January 17, 2018. https://towardsdatascience.com/random-forest-in-python-24d0893d51c0.

Lei, Deren, Zichen Sun, Yijun Xiao, and William Yang Wang. “Implicit Regularization of Stochastic Gradient Descent in Natural Language Processing: Observations and Implications.” arXiv.org, November 1, 2018. https://arxiv.org/abs/1811.00659.

Macukow, Bohdan. “Neural Networks – State of Art, Brief History, Basic Models and Architecture.” SpringerLink, September 9, 2016. https://link.springer.com/chapter/10.1007/978-3-319-45378-1_1.

Mishra, Sushruta, Hrudaya Kumar Tripathy, Pradeep Kumar Mallick, Akash Kumar Bhoi, and Paolo Barsocchi. “EAGA-MLP-an Enhanced and Adaptive Hybrid Classification Model for Diabetes Diagnosis.” Sensors (Basel, Switzerland), July 20, 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7411768

Pham, Trang, Truyen Tran, Dinh Phung, and Svetha Venkatesh. “Predicting Healthcare Trajectories from Medical Records: A Deep Learning Approach.” Journal of Biomedical Informatics, April 12, 2017. https://www.sciencedirect.com/science/article/pii/S1532046417300710.

Pupale, Rushikesh. “Support Vector Machines(SVM) — An Overview.” Medium, June 11, 2018. https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989.

Qin, Yifan, Jinlong Wu, Wen Xiao, Kun Wang, Anbing Huang, Bowen Liu, Jingxuan Yu, Chuhao Li, Fengyu Fu, and Zhanbing Ren. “Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type.” International journal of environmental research and public health, November 15, 2022. https://pubmed.ncbi.nlm.nih.gov/36429751/.

Richards, Selena E, Chandana Wijeweera, and Albert Wijeweera. “Lifestyle and Socioeconomic Determinants of Diabetes: Evidence from Country-Level Data.” PloS one, July 28, 2022. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9333224/.

Schmittdiel, Julie A, Wendy T Dyer, Cassondra J Marshall, and Roberta Bivins. “Using Neighborhood-Level Census Data to Predict Diabetes Progression in Patients with Laboratory-Defined Prediabetes.” The Permanente journal, October 5, 2018. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6175602/.

Sivasankari, S S, J Surendiran, N Yuvaraj, M Ramkumar, C.N Ravi, and R.G Vidhya. “Classification of Diabetes Using Multilayer Perceptron.” IEEE Xplore, April 23, 2022. https://ieeexplore.ieee.org/abstract/document/9793085/.

Temurtas, Hasan, Nejat Yumusak, and Feyzullah Temurtas. “A Comparative Study on Diabetes Disease Diagnosis Using Neural Networks.” Expert Systems with Applications, May 30, 2009. https://www.sciencedirect.com/science/article/abs/pii/S0957417408007306.

Verma, Nilesh. “XGBoost Algorithm Explained in Less than 5 Minutes.” Medium, September 7, 2022. https://medium.com/@techynilesh/xgboost-algorithm-explained-in-less-than-5-minutes-b561dcc1ccee.

Appendix A

UCI Dataset Features

The UCI dataset has 768 samples with the following features:

1. Number of times pregnant
2. Plasma glucose concentration a 2 h in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-hour serum insulin (mu U/ml)
6. Body mass index (kg/m^2)
7. Diabetes pedigree function
8. Age (years)

Appendix B

Factors used in Predicting Diabetes in the study Salameh et al. (2021)

In the study titled, *Prediction of diabetes and hypertension using multi-layer perceptron neural networks* the following factors were used to predict diabetes:

- Age
- Weight
- Fat-Ratio
- Glucose
- Insulin

Appendix C

Age-Wise Distribution of Diabetes: Percentage of the Presence of Diabetes Cases Across All Ages

Age Buckets	0; No Diabetes	1; Diabetes
(1 18]	40%	3%
(18 25]	8%	1%
(25 30]	5%	1%
(30 35]	5%	2%
(35 40]	5%	3%
(40 45]	5%	5%
(45 50]	5%	7%
(50 55]	5%	9%
(55 60]	5%	12%
(60 65]	5%	16%
(65 70]	4%	14%
(70 75]	3%	11%
(75 80]	5%	17%

Appendix D

Missing Data Distribution by Age Buckets: Percentage of Missing Data and Observations for Diabetes Across All Ages

Age_Buckets	Obs Missing	% Obs Missing
(1, 18]	590197	39%
(18, 25]	77506	5%
(25, 30]	52111	3%
(30, 35]	49070	3%
(35, 40]	47016	3%
(40, 45]	49296	3%
(45, 50]	46035	3%
(50, 55]	52880	4%
(55, 60]	53219	4%
(60, 65]	55549	4%
(65, 70]	45819	3%
(70, 75]	36433	2%
(75, 80]	82730	6%

Appendix E

Overview of Algorithms Used

SVM is a supervised machine learning model that is used for classification and regression. The objective of an SVM is to assign data points to one of the predefined categories within the dataset and make a continuous value prediction based on the input variables in the data (Pupale 2018). The algorithms create a hyperplane that separates the data into different classes. The algorithm aims to create a decision boundary that maximizes the separation between the two classes. An advantage of SVM models is that they work well with linear data and non-linear data. According to Pisner and Schnyer (2020), SVMmodels perform well with data with smaller sample sizes due to their simplicity and flexibility in classification problems.

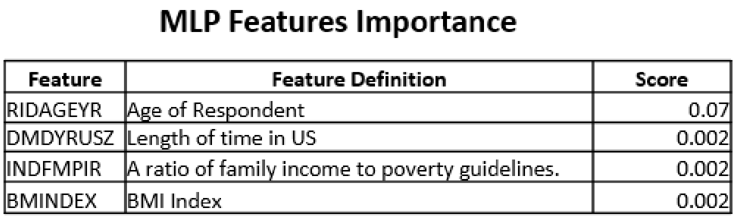
A RF model is a supervised learning model that is used for classification and regression. The models combine the output of multiple decision trees to return an output to reach a single result (IBM). It uses both bagging and random features to create an uncorrelated forest of trees. Unlike decision trees, RFs only select a subset of features in the dataset and are constructed using bootstrap sampling. These models also use bagging to add more variation and reduce the collinearity between trees. The primary hyperparameters in RF models are the number of trees, number of variables, and number of trees. The advantages of RF models are that they are able to effectively deal with missing values due to bagging, it is easy to find significant features by using Gini feature importance, and they reduce overfitting due to the vast number of decision trees.

XGB is a decision tree ensemble that uses gradient boosting where the errors in the model are minimized through gradient descent (Morde 2019). These models minimize a regularized L1 and L2 objective function that combines a convex loss function and a penalty term for model complexity. Each tree in the model is trained on a subset of the dataset and the predictions from each tree are used to make a final decision for each classification task (Verma 2022). These models handle missing data well, prevent overfitting due to its regularization techniques, and can handle many different types of data.

MLP models are feed forward neural networks that include input, hidden, and output layers. The input layer takes the signal or data to be handled, and the output layer is the classification layer for prediction. Since MLP models are feed forward networks, these models flow data forward to the output layer and the neurons in the model are trained using back propagation (Abirami and Chitra 2019). Backpropagation can fix issues that are not linearly separable and are structured to the approximation of every continuous function (Abinaya and Devi, 2022). These models are effective in identifying linear and non-linear relationships, identifying relationships between disparate datasets, predicts and missing data well, and can avoid overfitting due to weight decay and drop out. We believe that these features make MLP models effective in predicting diabetes with our complex dataset.

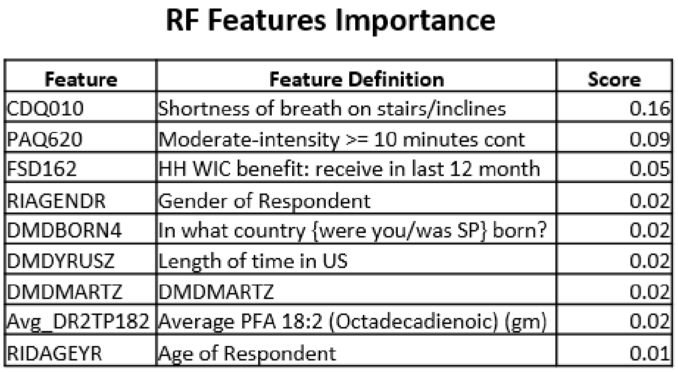
Appendix F

Model Feature Importance



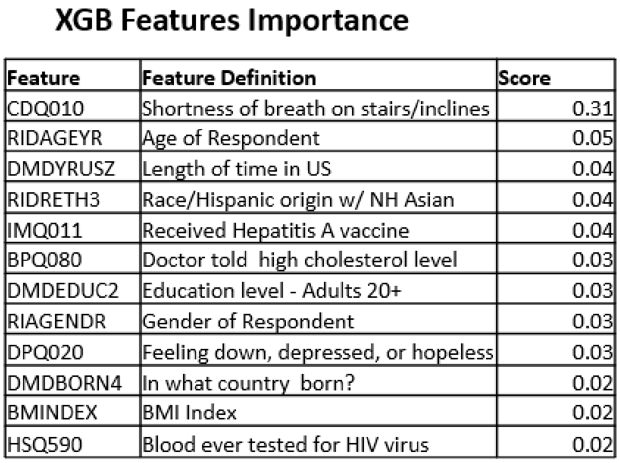
MLP Features Importance

Feature	Feature Definition	Score
RIDAGEYR	Age of Respondent	0.07
DMDYRUSZ	Length of time in US	0.002
INDFMPIR	A ratio of family income to poverty guidelines.	0.002
BMINDEX	BMI Index	0.002



RF Features Importance

Feature	Feature Definition	Score
CDQ010	Shortness of breath on stairs/inclines	0.16
PAQ620	Moderate-intensity >= 10 minutes cont	0.09
FSD162	HH WIC benefit: receive in last 12 month	0.05
RIAGENDR	Gender of Respondent	0.02
DMDBORN4	In what country {were you/was SP} born?	0.02
DMDYRUSZ	Length of time in US	0.02
DMDMARTZ	DMDMARTZ	0.02
Avg_DR2TP182	Average PFA 18:2 (Octadecadienoic) (gm)	0.02
RIDAGEYR	Age of Respondent	0.01



XGB Features Importance

Feature	Feature Definition	Score
CDQ010	Shortness of breath on stairs/inclines	0.31
RIDAGEYR	Age of Respondent	0.05
DMDYRUSZ	Length of time in US	0.04
RIDRETH3	Race/Hispanic origin w/ NH Asian	0.04
IMQ011	Received Hepatitis A vaccine	0.04
BPQ080	Doctor told high cholesterol level	0.03
DMDEDUC2	Education level - Adults 20+	0.03
RIAGENDR	Gender of Respondent	0.03
DPQ020	Feeling down, depressed, or hopeless	0.03
DMDBORN4	In what country born?	0.02
BMINDEX	BMI Index	0.02
HSQ590	Blood ever tested for HIV virus	0.02