Sugar & Spice: Shallow Machine Learning vs Deep Learning for Diabetes Classification

William Reindl, Raja Allmdar Tariq Ali, Nikhil Dhoka, Aditya Pawar
Dept. of Computer and Information Science
Indiana University Purdue University Indianapolis
Indianapolis, Indiana 46202

***Abstract*—The common machine learning method used in medical classification tasks today revolves around shallow networks and so coined “classical” statistical methods. Diabetes prediction is a difficult problem due to multiple interconnected factors resulting in “less than ideal”. Classical methods have excellent performance for low computational time with comparable medical papers, placing AUC performance of about 0.90. We compare the performance of deep neural networks like MLP (Multi Layered Perceptron) [1] and classical methods like Random Forest [2]. We also demonstrate how deep neural networks like, MLP can handle significant imbalances in diabetic and non-diabetic patients.**

***Index Terms*—AI, classification, multi layered perceptron, diabetes**

I. Introduction

Deep learning models have had wide spread use in the image processing field and anomaly detection for time series data using LTSM networks. Deep learning models unfortunately require large amounts of data to find solutions to problems. The more complicated the problem the more layers you need to find different ways to separate the data, which requires more data. Classical statistical methods require less data since they do not need to find the problem but iterate over it. Advances in healthcare monitoring systems like smart watches has enabled more data to be available. Multi Layered Perceptrons are a easy solution that increases AUC significantly vs classical methods.

II. Methods

For our testing we used the Diabetes prediction dataset [3], which uses real medical data compiled from over 95997 real samples. Our ratio of non-diabetic to diabetic people is skewed heavily to the former which represents 92% of our total data. We have 8028 samples of people with diabetes and 87968 samples of people without diabetes. To control for this heavy in balance in our data, we randomly sample 8028 samples out of the 87968 people without diabetes, thus making it 16056 total of a 50:50 ratio. Train, Validation and Test sets are split 80:10:10 respectively with randomness between splits controlled by a fixed constant.

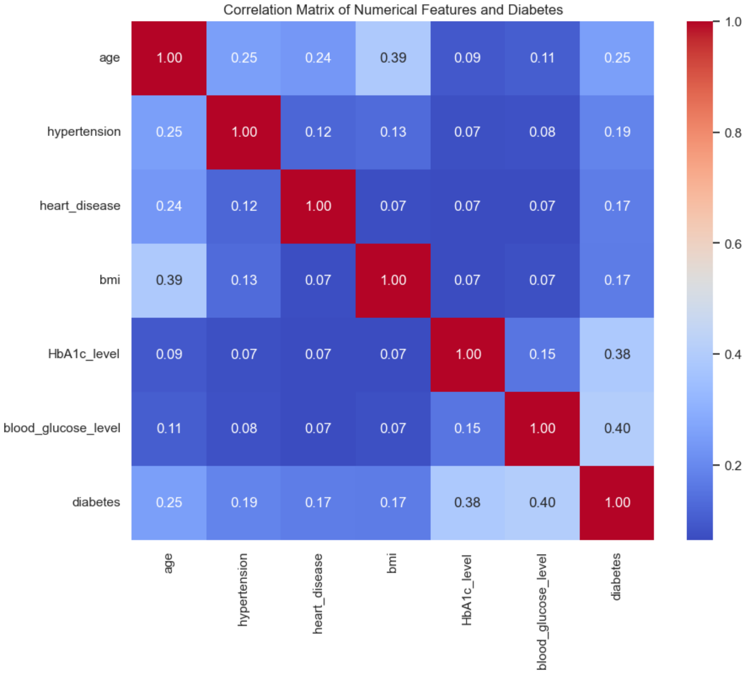

Fig. 1. Correlation matrix showing the relation between features.

In Fig. 1 we see the relationship features in the dataset, where multiple variables have a strong covariance with either that can influence each others values. HbA1c and blood glucose levels had the strongest correlation between each other this makes sense intuitively since these are direct measurements for diabetes and elevated levels in these areas will nearly guarantee a diabetes diagnosis. Heart disease or higher BMI would seem to indicate in this matrix that it matters very little in diagnosis but both can contribute to Diabetes.

A. MLP

The PyTorch library was used to manually create the Multi Layered Perceptron [4]. All models weights were initialized as zeroes and were run for 200 epochs each at a batch size of 16 samples per batch. This size was picked due to a decrease in Precision and Recall scores when upping to values of 32-64 samples per batch.

The MLP model uses 4 fully connected layers with ReLU as their activation function. The sigmoid layer at the ends binds the outputted tensor values between a [0, 1] using the following formula

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}} = 1 - \sigma(-x). \qquad (1)$$

The number of neurons used in each layered was determined by combining the number of input and output features and taking the mean of them 2.

$$NumNeurons = \frac{Input\ Features + Output\ Features}{2} \tag{2}$$

For our optimization function we used Adam with a constant learning rate of 1e-3. Our loss function used for back propagation was BCE(binary cross entropy) due to the single class used for prediction.

B. Random Forest

We used the sckit-learn [5] random forest ensemble implementation for our training/testing. We initialize a blank unpopulated tree with our randomness fixed between runs. We use the Gini tree algorthim to determine if the trees nodes should be split further. The batch size is 2 and no restriction is placed on the max depth of tree.

III. Results

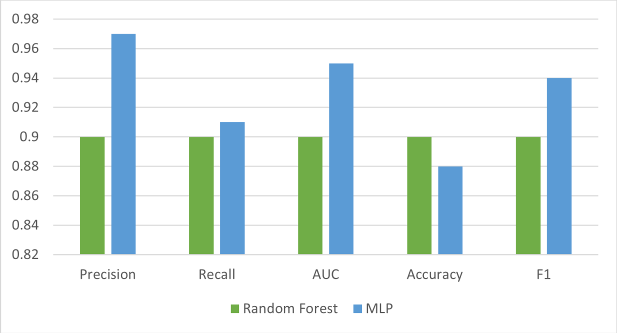

Fig. 2. Performance metrics for models.

Random Forest performed the worst out of the 2 models with every single metric being lower then MLP except for Accuracy. This means that MLP was over predicting patients as diabetic but Random Forest was more hesitant to assign this class as can be seen in the recall metric. Precision and Recall scores 0.900692 and 0.900685 respectively. The AUC (Area Under the Curve), Accuracy and F1 was 0.900691, 0.900685 and 0.900686 in that order. Performance was very consistent between metrics, deviation by 0.00001 points.

The Multi Layered Perceptron preformed the best in comparison to random forest. Precision and Recall scored 0.965812 and 0.91129 respectively. AUC, Accuracy and F1 scored 0.953985, 0.879826 and 0.937759 in that order. MLP achieved a 6.25% improvement over random forest at its max and recording its lowest improvement at 4.05%. Accuracy did decrease by 3.44% for the Multi Layered Perceptron.

IV. Conclusion

Multi Layered Perceptions can improve overall scores dramatically over traditional well accepted methods such as random forest [6]. This conclusion was replicated by another paper [7] with their newest MLP augmentation achieving a 0.95 in AUC. This paper came to the same result as we did in which classical methods preformed statistically worse then their deep learning counterparts. While not a replication study, we believe our work draws similar parallels. Additionally, our dataset is x21 times larger then their Pima Indians Diabetes Database [8], which robustly corroborates their results.

Performance did decrease for precision in the MLP deep network however, more emphasis on over predicting diabetes might be more desirable then under reporting. Further work will be required to determine if accuracy can be improved for the model without harming recall.

V. Future Work

Further testing of the dataset is needed to verify the training and testing is not biased to a specific class such as race or gender. Men might be more prone to diabetes due to other external factors and models could then be biased to converge using that factor. A wider range of models would be used to test that the pit falls of a particular style of learning (classical vs deep learning) are reproducible regardless of an architectures implementation. Examples would be ResNet-16 for more deep learning models and Naive Bayes classification algorithm for the classical.

Further research beyond the scope of this paper would call for deep learning roles in medical time series data, such as cancer growth over a fixed time. LTSM and convolution networks can be used to "remember" previous information and bias current observations to a particular class.

Acknowledgment

The authors would like to acknowledge Edwin Sanchez for proof reading the paper.

References

[1] S. Haykin, *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[2] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[3] M. Mustafa, "Diabetes prediction dataset," Apr. 2023. [Online]. Available: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[4] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," *CoRR*, vol. abs/1912.01703, 2019. [Online]. Available: http://arxiv.org/abs/1912.01703

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] W. Xu, J. Zhang, Q. Zhang, and X. Wei, "Risk prediction of type ii diabetes based on random forest model," in *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, Feb 2017, pp. 382–386.

[7] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76 516–76 531, 2020.

[8] V. Sigillito, "Pima indians diabetes database," Apr. 2017. [Online]. Available: https://data.world/uci/pima-indians-diabetes