Bayesian Neural Networks vs Artificial Neural Networks

Sarah Beaver and Mica Haney
University of North Texas

Artificial neural networks, whatever the type, are the most prominent and most basic of deep learning models. This group of well-known models has a variant known as Bayesian neural networks, which combine artificial neural networks with Bayes' Theorem in an attempt to achieve better learning by looking at the dataset more often and more closely than in a feedforward, fully-connected network. This project seeks to evaluate if Bayesian neural networks are comparable to other artificial neural networks in an attempt to look at if Bayesian neural networks should see wider use than they currently do.

1. Introduction

Since the introduction of the perceptron [8], artificial neural networks have become the backbone of machine learning. The artificial neural network is reliable and appears in a large portion of machine learning models and practically defined deep learning. Bayesian neural networks [2, 7] combine artificial neural networks with Bayes' Theorem. The effect of this is that a Bayesian neural network not only generates predictions like an artificial neural network, but also a confidence interval. Unlike probabilistic predictions made by artificial neural networks, a Bayesian confidence interval takes into account portions of the data seen more recently.

Given that Bayesian neural networks appear to be very close to an artificial neural network, it begs the question of if performance suffers or if two models of comparable construction are also comparable in performance. If not, it would explain why Bayesian neural networks are not more widespread. If so, then it would indicate that awareness should be raised so that Bayesian neural networks can be investigated and developed further.

This paper seeks to compare artificial and Bayesian neural networks to investigate the potential of Bayesian neural networks as a replacement to artificial neural networks in appropriate applications.

2. Related Work

Work comparing artificial and Bayesian neural networks could not be found in the time spent researching the topic. However, research on applying Bayesian neural networks to disease prediction tasks exists in abundance. Since the experiments were run on data from All of Us [1] with the intention of predicting disease diagnosis, a selection of papers is mentioned.

[6] looks at applying a Bayesian neural network to predict heart disease. The authors of [9] try to predict cancer and diabetes occurrences using Bayesian neural networks. In [4] not only is a Bayesian neural network developed, but also a benchmark built upon the Kaggle Diabetic Retinopathy Detection Challenge [3].

3. Methods

While individual models were tuned for the hyperparameter selections that worked best for the model in question, both models were built in a fashion as to keep them relatively equivalent in size and complexity.

3.1 Data Preparation

Data was collected from the All of Us data repository [1]. All of Us collects information from participants including but not limited to electronic health records and surveys.

First we had to query for the three tables that we were interested in. The first table was a survey table which includes questions that the patient has answered from surveys. We selected the person id, the date the patient answered the question, the survey the question was from, the question, and the answer where the survey was from the survey of basics, lifestyle, overall health, personal medical history, and family medical history and the patient had a condition of Rheumatoid Arthritis, Sleep Apnea, Type 2 Diabetes Mellitus with hyperglycemia, or Acute Myocardial Infarction (Heart Attack). The second table query we selected the patients id, gender, date of birth, race, ethnicity, and sex at birth with the same condition requirement used in survey table query. The last table to be queried is the condition table where we selected the person id, condition id, condition date, condition name, and condition code where the code matched to one of the four condition of Rheumatoid Arthritis, Sleep Apnea, Type 2 Diabetes Mellitus with hyperglycemia, or Acute Myocardial Infarction (Heart Attack).

From three different selected tables from All of Us Research, the survey table included questions that included the patient's work, home, use of tobacco, alcohol, and recreational drugs, general health, daily activities, ect. Some examples of questions that we used were "Alcohol: Alcohol Participant", "Living Situation: How Many People", ect. The survey table is in the format of person id, question name, survey group, date. We included a table that contains the conditions that each patient had. This table was also in a similar format of patient id, condition, and condition date.The last table we included was patient demographics which included gender, date of birth, and race. Created a list of questions and a list of conditions. We looped through all the patient ids and used them to select the row from the patient demographic information, looped through each question and selected all the rows from the surveys table to select all the questions that the patient answered, and looped through each of the question and selected all the conditions that the patient has. We saved this as a separate file so that we could go back and access it at any time.

After verifying that there were no duplicate rows in the data and that it was in the format we needed, we started data cleaning.We removed all the dates from the data, patient id, and date of birth. We replaced answers such as "Prefer not to answer" and "Skip" with na. Any column that had more than five percent of missing data was dropped. This removed all the questions from the family history surveys such as "Diagnosed Health Condition: Father Circulatory Condition" and "Disability: Dressing Bathing" as well as some others. After reviewing correlation we found that the question "Biological Sex At Birth: Sex At Birth" from survey table and "Sex At Birth" column from patient demographic table had identical answers. Additionally "Sex At Birth" column and "Gender" column from patient demographic table had a 87% correlation. Therefore, we dropped the question "Biological Sex At Birth: Sex At Birth" column and "Sex At Birth" column leaving the "Gender" column to be used in analysis. We performed one hot encoding on the data and separated it into input section and target sections. Once more we went through the one hot encoding input section and removed any "na" columns that were generated. Below in Figure 1 shows the columns left before one hot encoding with how many "na" values and Figure 2 shows the input data columns after one hot encoding processing without "na" columns. Also Figure 2 shows the target column names.

```
gender                                                    265
race                                                      312
ethnicity                                                 312
Alcohol: Alcohol Participant                              300
Cigar Smoking: Cigar Smoke Participant                    403
Education Level: Highest Grade                            329
Electronic Smoking: Electric Smoke Participant            339
Employment: Employment Status                             318
Home Own: Current Home Own                                470
Hookah Smoking: Hookah Smoke Participant                  353
Living Situation: How Many Living Years                   197
Living Situation: How Many People                         502
Living Situation: Stable House Concern                    230
Marital Status: Current Marital Status                    331
Overall Health: Difficult Understand Info                 401
Overall Health: Everyday Activities                       283
Overall Health: General Mental Health                     283
Overall Health: General Physical Health                   375
Overall Health: General Social                            345
Overall Health: Health Material Assistance                307
Overall Health: Organ Transplant                          272
Overall Health: Social Satisfaction                       282
Smokeless Tobacco: Smokeless Tobacco Participant          347
Smoking: 100 Cigs Lifetime                                248
Obstructive sleep apnea (adult) (pediatric)                 0
Type 2 diabetes mellitus with hyperglycemia                 0
Acute myocardial infarction, unspecified                    0
Rheumatoid arthritis with rheumatoid factor, unspecified    0
```

Figure 1: Counts of "na" in columns before encoding.

```
['Female' 'Male']
['Another single population' 'Asian' 'Black or African American'
 'More than one population' 'White']
['Hispanic or Latino' 'Not Hispanic or Latino']
['Alcohol Participant: No' 'Alcohol Participant: Yes']
['Cigar Smoke Participant: No' 'Cigar Smoke Participant: Yes']
['College graduate or advanced degree'
 'Highest Grade: College One to Three' 'Highest Grade: Twelve Or GED'
 'Less than a high school degree or equivalent']
['Electric Smoke Participant: No' 'Electric Smoke Participant: Yes']
['Employed for wages or self-employed' 'Not currently employed for wages']
['Current Home Own: Other Arrangement' 'Current Home Own: Own'
 'Current Home Own: Rent']
['Hookah Smoke Participant: No' 'Hookah Smoke Participant: Yes']
['How Many Living Years: 1 to 2' 'How Many Living Years: 11 to 20'
 'How Many Living Years: 3 to 5' 'How Many Living Years: 6 to 10'
 'How Many Living Years: less 1' 'How Many Living Years: more 20']
['0' '1' '10' '2' '3' '4' '5' '6' '7' '8' '9']
['Stable House Concern: No' 'Stable House Concern: Yes']
['Current Marital Status: Divorced'
 'Current Marital Status: Living With Partner'
 'Current Marital Status: Married' 'Current Marital Status: Never Married'
 'Current Marital Status: Separated' 'Current Marital Status: Widowed']
['Difficult Understand Info: Always' 'Difficult Understand Info: Never'
 'Difficult Understand Info: Occasionally'
 'Difficult Understand Info: Often' 'Difficult Understand Info: Sometimes']
['Everyday Activities: A Little' 'Everyday Activities: Completely'
 'Everyday Activities: Moderately' 'Everyday Activities: Mostly'
 'Everyday Activities: Not At All']
['General Mental Health: Excellent' 'General Mental Health: Excllent'
 'General Mental Health: Fair' 'General Mental Health: Good'
 'General Mental Health: Poor' 'General Mental Health: Very Good']
['General Physical Health: Excellent' 'General Physical Health: Fair'
 'General Physical Health: Good' 'General Physical Health: Poor'
 'General Physical Health: Very Good']
['General Social: Excellent' 'General Social: Fair' 'General Social: Good'
 'General Social: Poor' 'General Social: Very Good']
['Health Material Assistance: Always' 'Health Material Assistance: Never'
 'Health Material Assistance: Occasionally'
 'Health Material Assistance: Often'
 'Health Material Assistance: Sometimes']
['Organ Transplant: No' 'Organ Transplant: Yes']
['Social Satisfaction: Excellent' 'Social Satisfaction: Fair'
 'Social Satisfaction: Good' 'Social Satisfaction: Poor'
 'Social Satisfaction: Very Good']
['Smokeless Tobacco Participant: No' 'Smokeless Tobacco Participant: Yes']
['100 Cigs Lifetime: No' '100 Cigs Lifetime: Yes']
```

Figure 2: Input columns (sample and label) after encoding.

The data was split using scikit-learn's train_test_split method with 56% for test, 24% for validation, and 20% for testing.

3.2 Neural Network Models

We trained a model to output all four disease outputs. It was tuned on having normalization or not, with number of layers 1-3, including a dropout or not, dropout rate, activation for layers, activation for output layer, and learning rate. The best model for all four disease targets did not have normalization. It had one hidden layer with 64 nodes, a dropout with a rate of .4, and activation function using relu. The activation function for the final layer was sigmoid and the learning rate was .00018. The resulting model can be seen in Figure 3 and results on the hyperparameter used for the best model can be seen in Table 1.

We performed the same testing with a single target for each disease. We made sure to select the target values so that the output was balanced. For Sleep Apnea there were 6630 patients with it and we pulled 6630 who did not. This gives a total of 13260 patients to test, validate, and train the model for Sleep Apnea. There were 6977 patients with Diabetes and we pulled another 6977 that did not have Diabetes. There were only 140 patients that had a Heart Attack and we pulled an extra 140 patients that did not have Heart Attack. There were only 58 patients that had Arthritis so we pulled those 58 and 58. For Heart Attack and Arthritis, the small sample made it hard to get valid testing results from the models.

The model for Sleep Apnea ended up having normalization. It had two hidden layers

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 93)                8742

 dense_1 (Dense)             (None, 64)                6016

 dropout (Dropout)           (None, 64)                0

 dense_2 (Dense)             (None, 4)                 260

=================================================================
Total params: 15,018
Trainable params: 15,018
Non-trainable params: 0
_________________________________________________________________
```

Figure 3: The best performing four disease prediction artificial neural network.

Hyperparameter	Sleep Apnea	Heart Attack Output	Arthritis	Diabetes	4 Disease Output
Normalization	True	False	False	False	False
Num. Layers	2	2	2	2	1
Units 0	224	384	448	352	64
Units 1	288	512	320	384	-
Units 2	-	-	-	-	-
Activation	tanh	relu	sigmoid	tanh	relu
Final Activation	sigmoid	sigmoid	sigmoid	sigmoid	tanh
Dropout	False	True	False	False	False
Drop Rate	-	0.2	-	-	-
LR	0.0012622944	0.0083145018	0.0006113303	0.0005353758	0.0001830801

Table 1: Artificial neural network hyperparameters after tuning.

Model	Dense Variational	DenseFlipout
Normalization	FALSE	TRUE
Num. Layers	2	1
Units 0	96	416
Units 1	384	-
Units 2	-	-
Activation	sigmoid	tanh
Final Activation	sigmoid	sigmoid
Dropout	TRUE	TRUE
Drop Rate	0.2	0.2
LR	0.0002502286688	0.00140170488

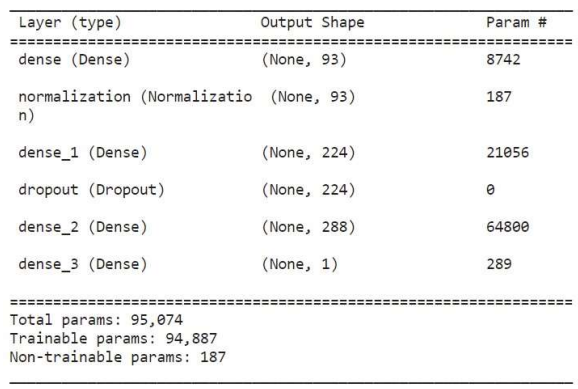
Table 2: Bayesian neural network hyperparameters after tuning.

with the first layer having 224 nodes with a dropout with a rate of .3 and the second layer had 288 nodes with no dropout using an activation of tanh on the hidden layers. The output layer used activation of sigmoid and the learning rate was .00126. The resulting model can be seen in Figure 4 and the results on the hyperparameter used for the model can be seen in Table 1.

The model for Diabetes did not use normalization. It had two hidden layers with the first layer having 352 nodes with a dropout rate of .4 and the second layer had 384 nodes with no dropout while using activation of tanh. The output layer used an activation of sigmoid and had a learning rate of .00053.

The model for Heart Attack did not use normalization. It had two hidden layers with the first layer having 384 nodes with no dropout and the second layer had 512 nodes with a dropout rate of .2 using an activation function of relu. The output layer had an activation function of sigmoid and a learning rate of .00831.

The model for Arthritis did not use normalization. It used 2 hidden layers with the first layer having 448 nodes with a dropout rate of .4 and the second layer had 64 nodes with no dropout using an activation function of sigmoid on the hidden layers. The output layer used sigmoid activation and had a learning rate of .0061.



Layer (type)	Output Shape	Param #
dense (Dense)	(None, 93)	8742
normalization (Normalization)	(None, 93)	187
dense_1 (Dense)	(None, 224)	21056
dropout (Dropout)	(None, 224)	0
dense_2 (Dense)	(None, 288)	64800
dense_3 (Dense)	(None, 1)	289

Total params: 95,074
Trainable params: 94,887
Non-trainable params: 187

Figure 4: Sleep apnea artificial neural network.

3.3 Bayesian Neural Network Models

We tested with two different types of layers for the Bayesian Neural Networks. The first is a DenseVariational layer and the second is a DenseFlipout layer. The DenseVariational layer [5] works very similarly to dense layers. Additionally it uses variational inferences to fit the posterior over the kernel and the bias. The DenseFlipout [5] layer assumes kernel and/or bias are drawn from distribution and does stochastic forward pass by sampling from the kernel and bias posteriors.

For both layers types we tried to get outputs for all four targets at the same time. We tuned for normalization, dropout, number of layers, number of units in each layer, activation function for each layer, activation function for the output layer, and the learning rate.

The model with the DenseVariational layers did not use normalization. It had two layers with the first layer having 96 nodes and the second layer having 384 nodes. These layers did have dropout with a value of .2 and activation functions of sigmoid. The output layer used an activation function of sigmoid. The learning rate on this model was .00025.

The model with DenseFlipout layers did use normalization. It had one layer with 416 nodes and an activation function of tanh. It had a dropout rate of .2. The activation function for the output layer was sigmoid. Its learning rate is .00140.

Class	Precision	Recall	F1
Sleep Apnea	0.48	1	0.65
Diabetes	0.51	1	0.67
Heart Attack	0	0	0
Arthritis	0	0	0

Table 3: Artificial neural network classwise results.

4. Results

For the artificial neural network results were taken for both validation and testing scores as described in Table 4. The model trained to predict sleep apnea achieved a 61.2% validation accuracy with a 61% testing accuracy. Diabetes has comparable results with 61.9% validation and 62% testing accuracy. The models trained for heart attack and arthritis got progressively more overfit. The heart attack model getting a 66.5% validation accuracy but a 57% testing accuracy while the arthritis model got a truly poor 67.1% validation accuracy and 33% testing accuracy. The poor scores for heart attack and arthritis predictions likely come from a poor dataset; the number of samples in the main dataset are small leaving only a very small subset of data to train the respective models on. The artificial neural network predicting all four diseases at once reported only a validation accuracy of 61.2%. In place of testing accuracy precision, recall, and F1 scores were calculated through both macro averaging of the classes as well as weighted averaging. The macro average F1 was a terrible 0.33, while the weighted average showed a better - but still poor - 0.65.

Due to issues with the All of Us workbench constantly crashing when results were being collected, only validation accuracy is reported for the Bayesian neural networks. The model using DenseVariational layers achieved an accuracy of 61.5%, while the model using

Model	Validation	Test						
			(macro avg)			(weighted avg)		
	(acc)	(acc)	(precision)	(recall)	(f1)	(precision)	(recall)	(f1)
Artificial Neural Network								
4 Disease Output	61.2		0.25	0.5	0.33	0.49	**0.99**	0.65
Sleep Apnea	61.2	61	**0.62**	0.61	0.61	**0.62**	0.61	0.61
Diabetes	61.9	**62**	**0.62**	**0.62**	**0.62**	**0.62**	0.62	**0.62**
Heart Attack	66.5	57	0.57	0.57	0.57	0.57	0.57	0.57
Arthritis	**67.1**	33	0.28	0.33	0.29	0.28	0.33	0.29
Bayesian Neural Network								
DenseVariational	61.5							
DenseFlipout	**62.8**							

Table 4: Results from all models. Of all of the artificial neural networks, the model predicting arthritis has the best validation accuracy while the one predicting diabetes has the best testing accuracy. Between all four disease prediction models, the Bayesian neural network with the DenseFlipout layers performs the best.

Dense Flipout layers achieved 62.8% accuracy (Table 4).

Some of the poor performance is due to the significant class imbalance in the dataset. With heart attack and arthritis labels accounting for such a small portion of the dataset, the four-disease models are heavily skewed towards predictions of sleep apnea and diabetes. Table 2 (taken from the four disease artificial neural network testing scores) demonstrates this with a complete inability to predict two of the four diseases.

5. Conclusion

Each model did almost as badly as possible. No model achieved more than 68% accuracy, and most did worse by barely achieving over 60%. Testing scores were often worse, some to the point of being no better than randomly guessing the diagnosis.

Both classes of models did equally poorly, achieving roughly equivalent results on their validation scores with the Bayesian network using DenseFlipout layer performing slightly better. However, given how poor the scores are it's impossible to say that the results indicate that the models are truly equivalent in performance. Therefore, it is impossible to conclude one way or the other if Bayesian neural networks should see more widespread use as a - possibly preferable - alternative to a traditional artificial neural network.

6. References

1. "All of Us Research Hub", Researchallofus.org, 2022. [Online]. Available: https://www.researchallofus.org/.
2. C. Bishop, "Bayesian Neural Networks", Journal of the Brazilian Computer Society, vol. 4, no. 1, pp. 61-68, 1997.

3. "Diabetic Retinopathy Detection | Kaggle", Kaggle.com, 2022. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection.
4. A. Filos et al., "A Systematic Comparison of Bayesian Deep Learning Robustness in Diabetic Retinopathy Tasks", 2022. Available: https://arxiv.org/abs/1912.10481.
5. "Module: tfp.layers | TensorFlow Probability", TensorFlow, 2022. [Online]. Available: https://www.tensorflow.org/probability/api_docs/python/tfp/layers.
6. M. Muibideen and R. Prasad, "A Fast Algorithm for Heart Disease Prediction using Bayesian Network Model", 2020. Available: https://arxiv.org/abs/2012.09429.
7. V. Mullachery, A. Khera and A. Husain, "Bayesian Neural Networks", 2018. Available: https://arxiv.org/abs/1801.07710.
8. F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain.", Psychological Review, vol. 65, no. 6, pp. 386-408, 1958. Available: 10.1037/h0042519.
9. P. Singh, "Better Application of Bayesian Deep Learning to Diagnose Disease", 2021. Available: https://ieeexplore.ieee.org/abstract/document/9418301.