Feature Engineering and Missing Value Imputation for Classification of Depression with Deep Neural Networks

CAROLINE SKLAVER

GWU DATA SCIENCE CAPSTONE

08/20/2020

Outline

- Background & Data
- Problem Statement
- Literature Review
- Methods
  - Data Pre-processing
  - Data Imputation
  - Preliminary Analyses
  - Deep Neural Networks
  - Feature Engineering
- Results
- Conclusions

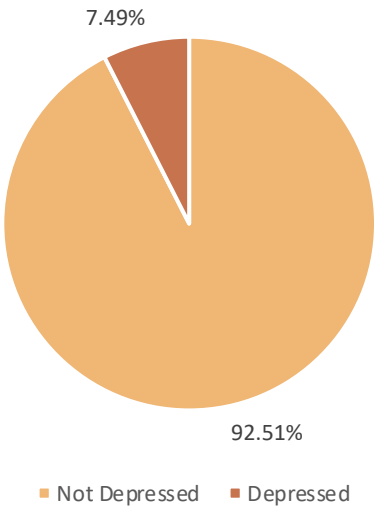
Background

- 1 in every 6 adults in the US will experience depression during their lifetime, which is about 16 million adults each year
- Depression is characterized as a long-term affect in mood caused by a complex combination of genetic, physiological, environmental, and social factors.
- Depression frequently goes undiagnosed and untreated, contributing to the overall burden of disease

National Health & Nutrition Examination Survey (NHANES)

Component	Demographics	Examination	Laboratory	Questionnaire	Dietary
Component Data Files	Demographics (including survey design variables)	Audiometry Blood Pressure Body Measures Muscular Strength Oral health Vision Exam etc...	Urine Collection Hepatitis HIV Heavy metals Plasma Glucose Total Cholesterol Triglycerides etc...	Alcohol use Balance Blood Pressure Diabetes Drug Use Social Support Vision Weight History etc...	Dietary Interview Supplement Use etc...

- Sample of the US population amounting to about 5,000 participants each two-year period
- Patient Health Questionnaire nine-item (PHQ-9) depression screening instrument assesses the frequency of depression symptoms over the past 2 weeks
  - Response categories include "not at all," "several days," "more than half the days," and "nearly every day" given a point ranging from 0 to 3. A total score of 10 or higher marked as '*depressed*'
- 31,357 complete depression observations between 2005-2016
- **7.49%** of the observations are classified as "*depressed*"



Features

Survey Component	Number of Features	Binary	Multi-class	Continuous
Demographics	7	1	2	4
Dietary	16	1	0	15
Examination	7	0	1	6
Laboratory	15	2	0	13
Questionnaire	67	43	14	10
Total	112	47	17	48

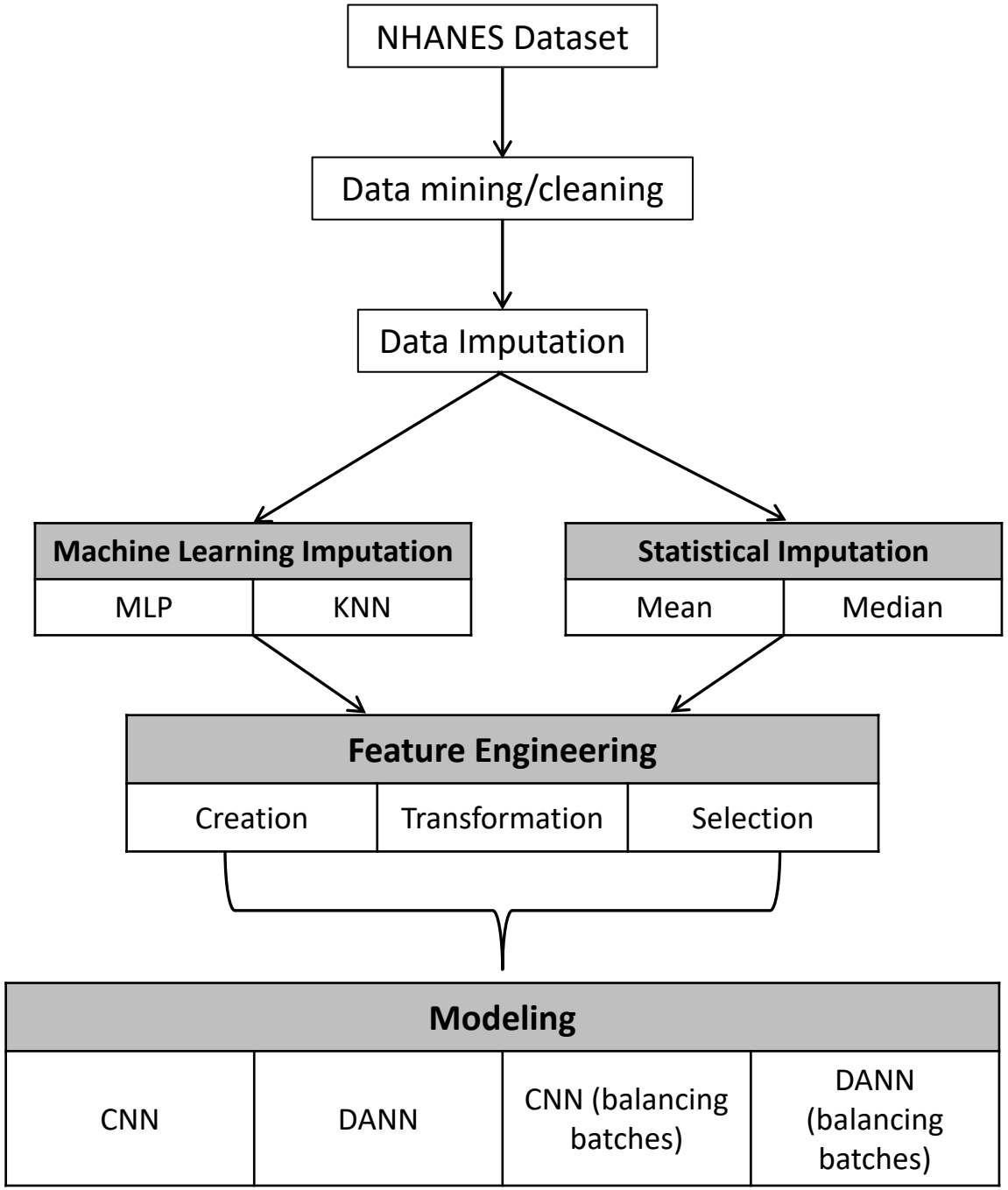
- Originally only 53 NHANES features imported (used in preliminary analyses)
- Features were chosen based on prior associations found in the literature
- Continuous features are utilized for transformations in feature engineering

Problem Statement

Depression is a widely undiagnosed mental health condition that greatly impacts the burden of disease. NHANES data is highly imbalanced (7.49% positive) and contains many missing values. The primary goal of this project is to identify the best imputation and feature engineering methods to optimize depression classification ability of a multi-layer artificial neural network (ANN) and a multi-layer convolutional neural network (CNN).

Literature Review

- Linear **associations with depression**:
  - Demographics such as age, race, gender, and income level [1]
  - Health characteristics including physical activity [2], metabolic syndrome, BMI [3], and other chronic conditions such as asthma, kidney disease [4], and osteoporosis
  - Dietary practices including caffeine consumption [5], household food insecurities [6], and Vitamin B levels [7]
- Jerez et al. 2020 [8] used **imputation methods** based on statistical techniques (mean, hot-deck and multiple imputation) and machine learning techniques (multi-layer perceptron (MLP), self-organization maps (SOM) and k-nearest neighbor (KNN))
  - Machine learning imputation performed best for ANN classification of breast cancer
- Heaton et. al. 2016 [9] reviewed **feature engineering** techniques for machine learning and found a DANN able to handle addition, summation and multiplication of features. Additionally, feature counts, differences, power transformations, and rational polynomial transformations were all synthesized by the DANN relatively easily
- **ANNs** have demonstrated accurate classifications of many diseases including heart disease [10], caries [11], and osteoporosis [12, 13]
- **Shallow CNN** to predict Coronary Heart Disease with NHANES data [14]
- **Support vector machines (SVM) and ensemble modeling** to predict risk of Diabetes with NHANES data [15]
- **Repeated random sub-sampling & Random Forests** (RF) to predict the risk of eight chronic diseases [16]


NHANES Dataset
Data mining/cleaning
Data Imputation
Machine Learning Imputation
MLP
KNN
Statistical Imputation
Mean
Median
Feature Engineering
Creation
Transformation
Selection
Modeling
CNN
DANN
CNN (balancing batches)
DANN (balancing batches)

Methods

Data Mining & Cleaning

- Data was downloaded and converted to CSV files using python script found on GitHub [17]
- Remaining preprocessing including importing files, extracting and renaming columns, and merging data for each year was done in Python Jupyter Notebook
- The target column of depression was calculated using the sum of the Patient Health Questionnaire (PHQ-9) nine-item survey ranging from 0 to 27, with a score of 10 or higher marked as '*depressed*'. Only data with complete PHQ-9 questionnaires were used in this project
- Responses of '*1 – yes*' and '*2 – no*' were converted to binary 1/0, respectively
- All responses of '*7 – don't know*' or '*9 – refused*' were marked as N/A
- Prior to the implementation of any algorithm predicting depression, all multi-class features are encoded, and continuous features are standardized to a mean of 0 and standard deviation of 1

Data Imputation

- Statistical methods:
  - Simple mean, median, and mode missing value imputation
  - Implemented after dropping columns with a proportion of missing values exceeding a given input threshold (0.75)
- Machine Learning methods:
  - Aim to use available information within the dataset to estimate and substitute the missing values
  - First, separated the data frame into complete data and incomplete data
  - Complete data (excluding the ultimate target variable '*depressed*') was used to predict values of the next-most-complete column. These predicted values were then imputed into that column and used with the complete data to predict the missing values of the subsequent column
  - Thus this method progressively predicts missing values with the imputed data
  - Used both Multi-layer perceptron (MLP) k-nearest neighbors (KNN) – regression or classification depending on target column

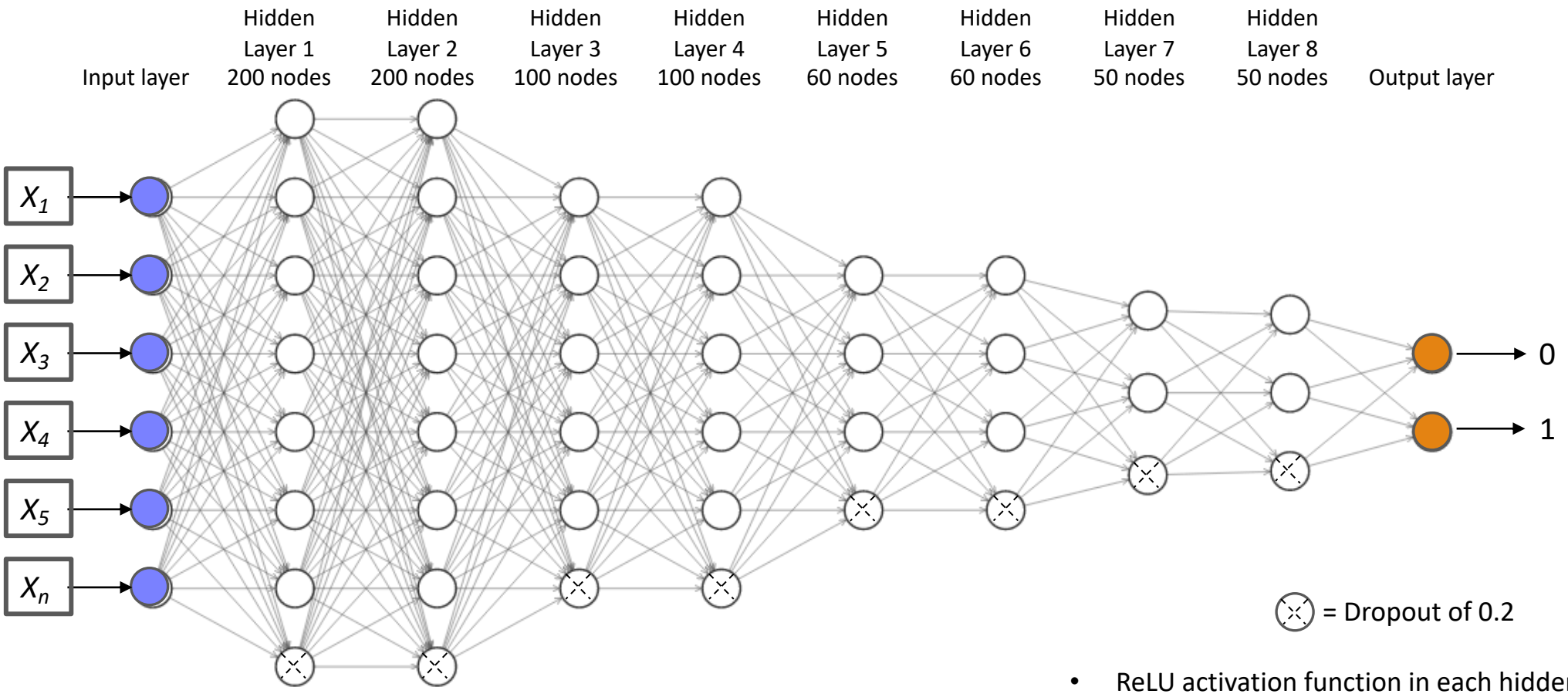
Preliminary Analyses: Results

- Original 53 columns of NHANES data
- With such low F1-scores, and accuracies averaging at 92%, the challenge with imbalanced data is highlighted in these results
- These results lead to the addition of features and construction of deep neural networks (DNNs)

F1-SCORES

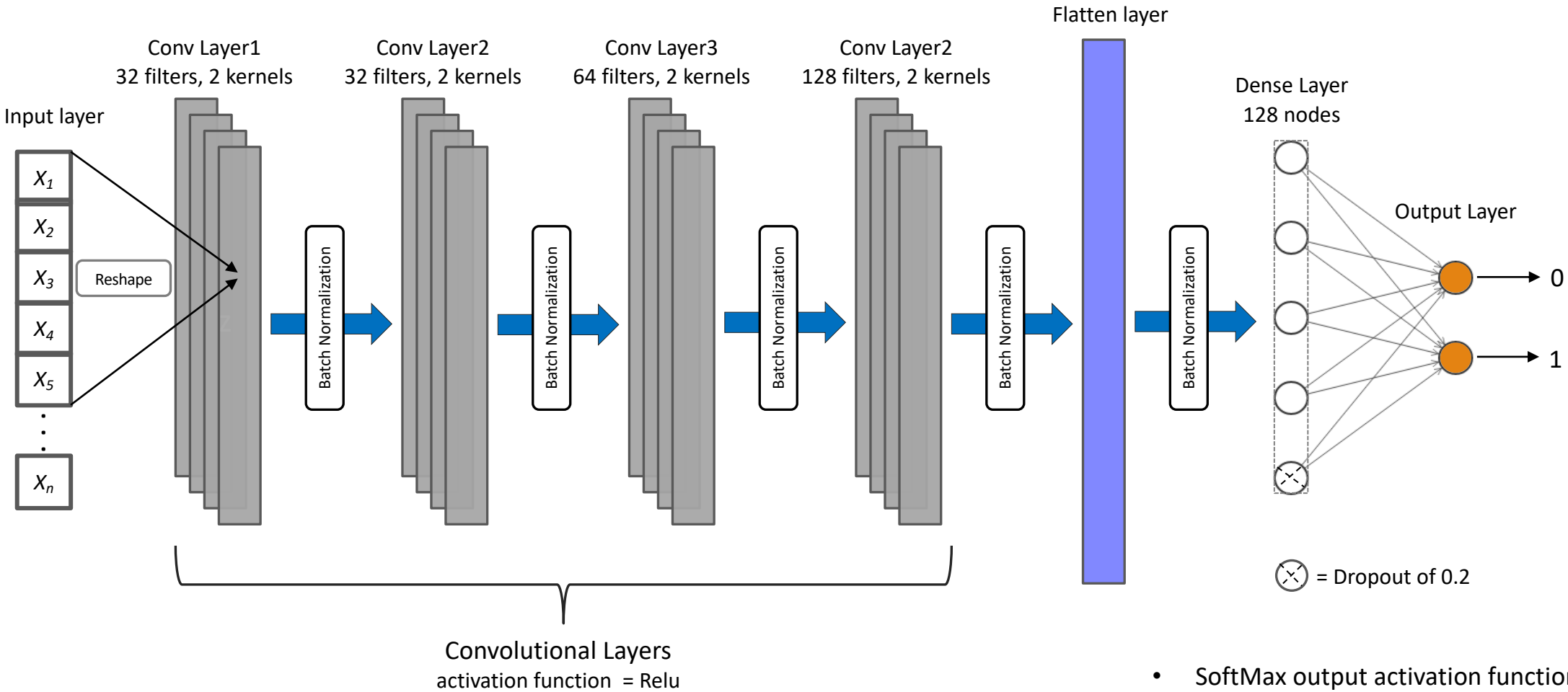
Imputation Type	Logistic Regression	Decision Tree	Random Forest	XGBoost	KNN	Naïve Bayes	Single Perceptron	MLP
MLP	0.1575	0.0129	0.0129	0.1839	0.0946	0.2465	0.2822	0.1489
KNN	0.1430	0.0077	0.0077	0.1630	0.0943	0.2459	0.0303	0.1675
Mean	0.1398	0.0154	0.0154	0.1833	0.0953	0.2473	0.1171	0.2346
Median	0.1398	0.0154	0.0154	0.1833	0.0953	0.2473	0.1171	0.0832

Deep Neural Network



- ReLU activation function in each hidden layer
- SoftMax output activation function
- Binary Cross Entropy loss function

1-D Convolutional Neural Network



- SoftMax output activation function
- Binary Cross Entropy Loss Function

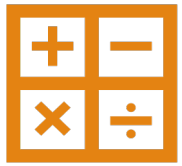
Verification of Neural Networks

- Implementation of the neural networks to classify:
  1. Different target variable within NHANES dataset – *'overweight'*
  2. Classification of *'diagnosis'* with UCI Breast Cancer dataset

			F1-SCORES	
Data	Target Variable	% positive cases	DANN	CNN
UCI Wisconsin Breast Cancer	*diagnosis*	37.26	0.8859	0.9385
NHANES MLP imputed	*overweight*	33.75	0.7143	0.7538

- Breast Cancer high F1-Scores confirm the architectures and implementations of the DNNs
- NHANES target of *'overweight'* F1-Scores are not as impressive
  - Features in the NHANES dataset have less predictive power than those in the Breast Cancer dataset
- Nevertheless, classification of '*overweight*' result in much higher F1-scores than in the case of depression
  - Difference in class imbalance proportions (33.75% vs. 7.49%)
  - The nature of the two variables, '*overweight*' having more recognizable attributes
- Conclusions:
  - DNNs are implemented and functioning properly
  - The high level of class imbalance with respect to depression is a significant hindrance
  - The NHANES data may, by nature, not be highly discriminatory or indicative for any classification

Feature Engineering



Creation

Addition of feature columns by either counting, summing, subtracting, dividing, or multiplying two or more existing features

15 new features created

Transformation

New features consisting of the square root, log, power, polynomial, and sine functions applied to the original continuous numerical features

Continual augmentation of transformed features to existing features

Selection

Sub-setting of features

Feature association with depression (literature review)

Reduce correlation between features

Feature Engineering Procedure	Number of Features (after encoding)
NHANES original	79
NHANES baseline features	166
Baseline + Created features (BCF)	186
BCF + log(cont)	240
BCF + $(cont)^{1/2}$	240
BCF + $(cont)^{2}$	240
BCF + polynomial	240
BCF + sin(cont)	240
BCF + $(cont)^{2}$ + log(cont)	293
BCF + $(cont)^{2}$ + log(cont) + $(cont)^{1/2}$	346
BCF + $(cont)^{2}$ + log(cont) + $(cont)^{1/2}$ + polynomial	399
BCF + $(cont)^{2}$ + log(cont) + $(cont)^{1/2}$ + polynomial + sin(cont)	452
Subset of CF	105
Subset + log(cont)	135
Subset + log(cont) + $(cont)^{1/2}$	165

Balancing Batches (BB)

- For each data imputation method and feature engineering experiment, both the DANN and CNN were trained with standard training data (imbalanced) and with balanced mini-batches
- Balanced mini-batches used random under sampling of the majority class without replacement to create a random, balanced batches of size 32
- Implemented to determine whether resampling methods are beneficial in these DNN classifications

Evaluation Metrics

Results are evaluated using F1-Score:

$$2 * \frac{Precision * Recall}{Precision + Recall}$$

Where,

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

	Predicted: 0	**Predicted: 1**
Actual: 0	TN	FP
Actual: 1	FN	TP

MODEL	AVERAGE F1-SCORES
DANN	0.1458
CNN	**0.2640**
DANN BB	0.2513
CNN BB	0.2488

Results: Model Comparison

- Average across all feature engineering techniques
- CNN has highest average

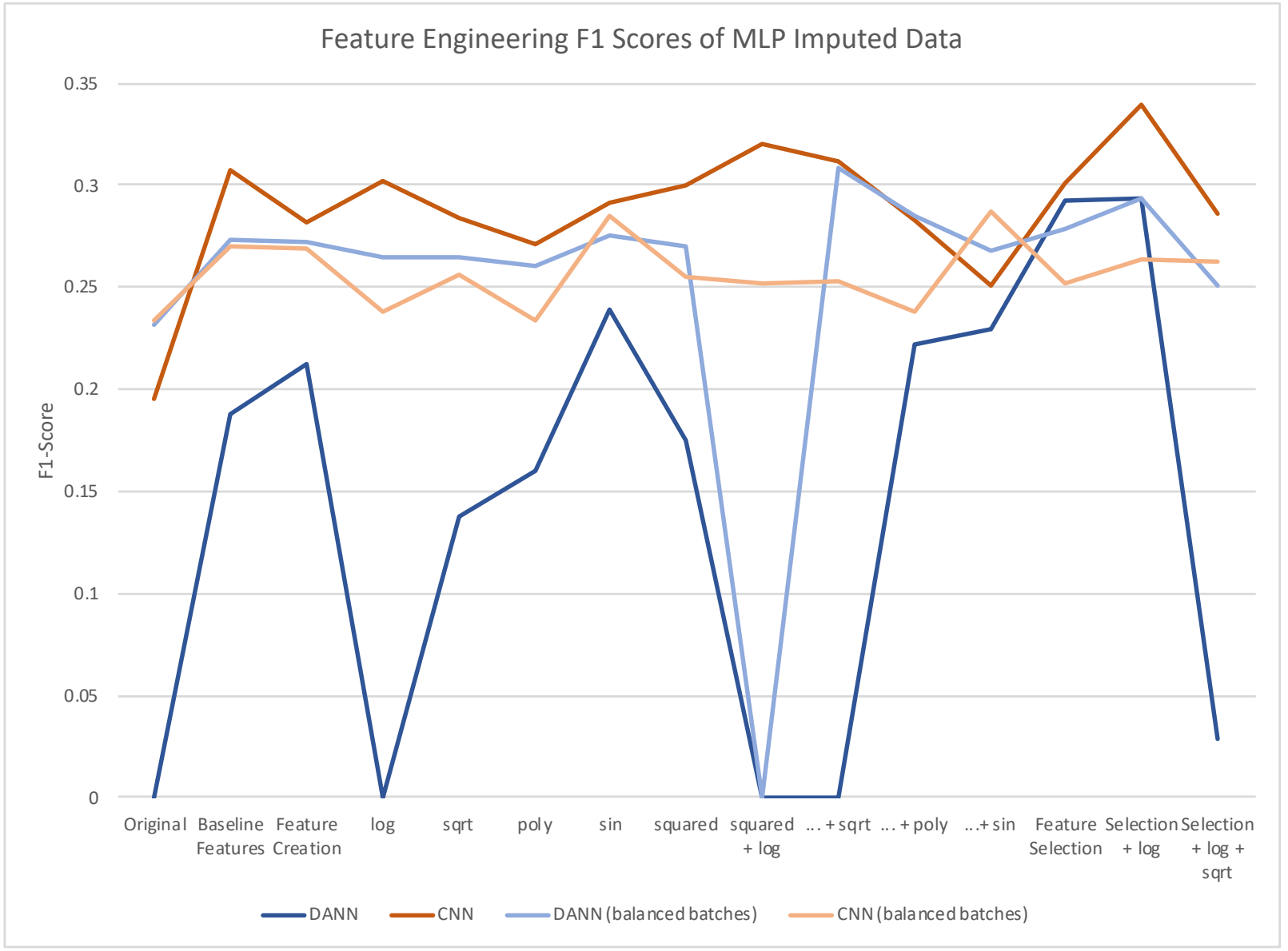
IMPUTATION METHOD	AVERAGE F1-SCORE (ALL MODELS)	CNN AVERAGE F1- SCORE
MLP	0.2312	**0.2794**
KNN	**0.2429**	0.2726
Mean	0.2083	0.2521
Median	0.2274	0.2520

Results: Imputation Methods

- Results averaged across all feature engineering results
- KNN provides the highest average F1-score averaging across all models
- Average F1-score of just the CNN results, MLP imputation has the best F1-score

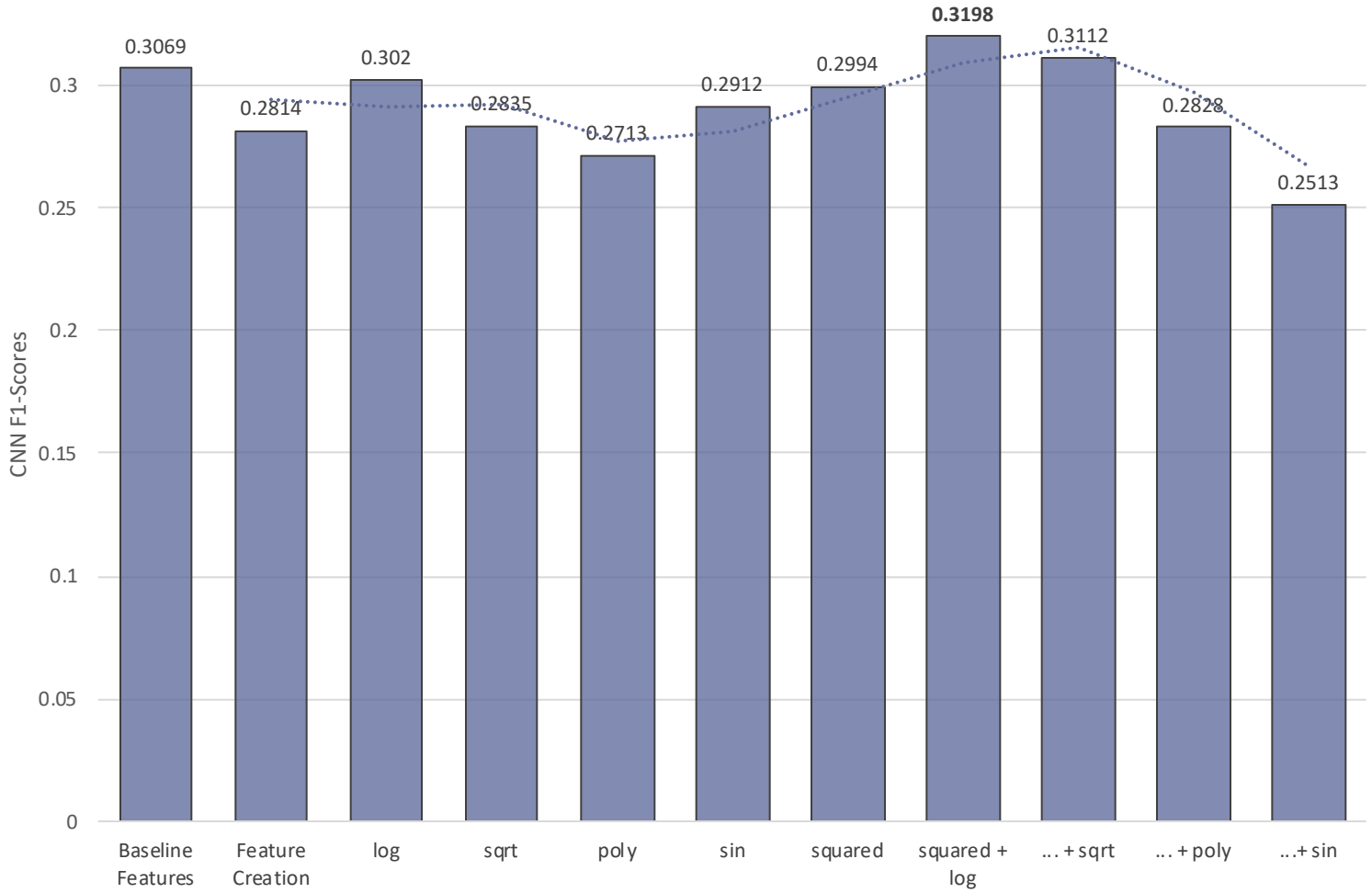
Results: Feature Engineering

- DANN has the greatest fluctuations and is the least reliable model
- DANN is generally improved by training on balanced batches.
- All models benefit from the addition of features from *'original'* to *'baseline features'*
- Created features may have no benefit
- Transformations have varying effects
- Most models benefit from some feature selection and subsequent transformations



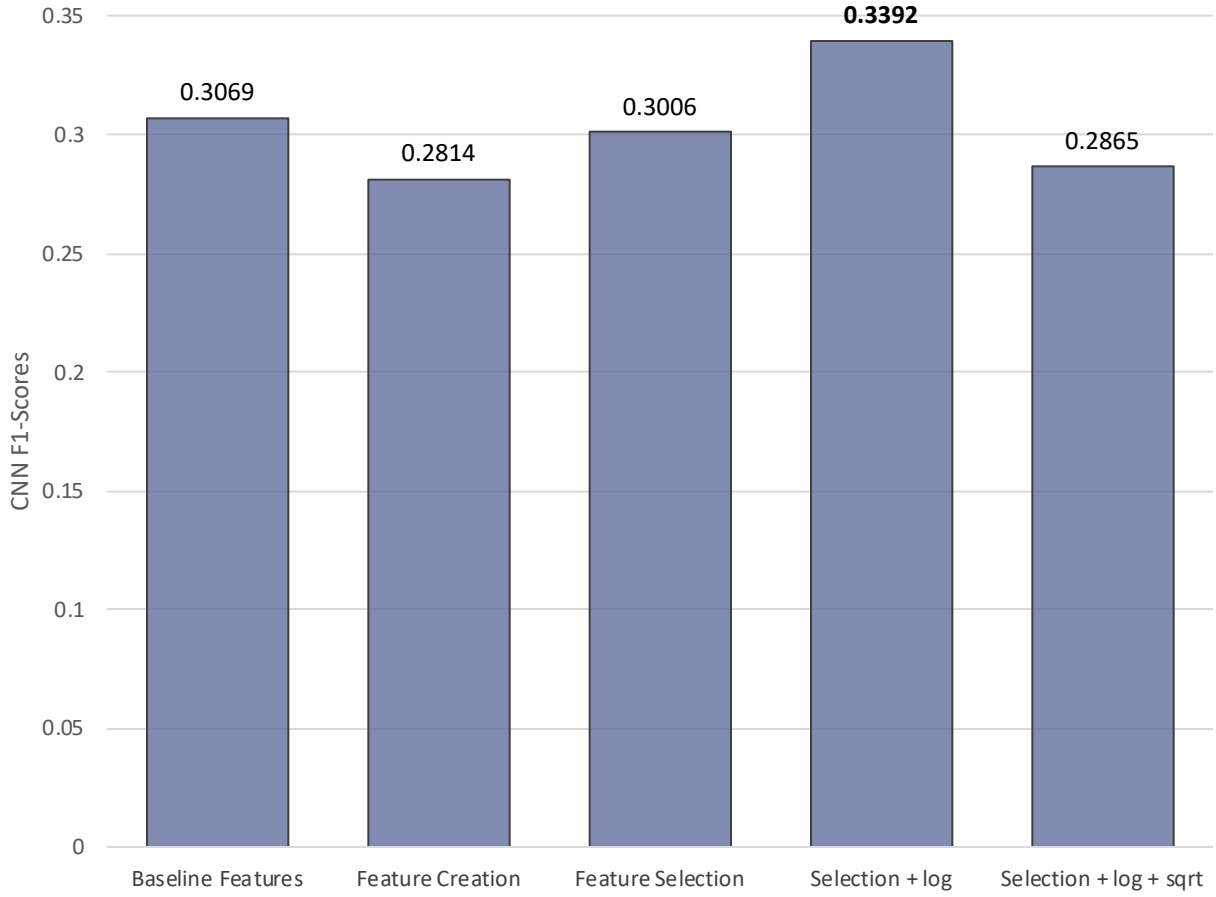
CNN and MLP Imputation

- Feature creation results in a decrease in the F1-score
- Log, sin, and squared transformations have better outcomes than polynomial and square root
- The best transformation and augmentation combination is the addition of both the square and log, resulting in a higher F1-score than the baseline (0.3198)
- Subsequent augmentations negatively affect performance.



CNN and MLP Imputation Cont.

- Feature selection alone leads to improvement from feature creation
- The addition of the log-transformed features to this subset provides the best results (0.3392) among all F1-scores of each imputation, model, and feature engineering combination.



Best Results

This model is achieved using the CNN without balancing batches, MLP imputation, and a subset of the extended number of NHANES features along with the 15 created features, and the log of continuous features

This F1-Score is an 11% increase from that without any feature engineering (0.3069)

	Predicted: 0	**Predicted: 1**
Actual: 0	9036	545
Actual: 1	499	268

F1-Score	0.3392
Precision	0.3296
Recall	0.3494
ROC AUC	0.6463
Accuracy	0.8991

Summary & Conclusions

- Depression is a complex mental disease with unspecified causes. Despite the unique combination of demographic, dietary, clinical, and questionnaire data from NHANES, I am unable to build a model to accurately classify depression
- The nature of this data with only 7.49% positive cases and many missing values did not provide enough information for DNNs to predict depression in the NHANES population.
- Progressive machine learning imputation out-performed standard statistical imputation techniques.
- The 1D CNN classification results were superior to the more commonly used machine learning algorithms
- Feature engineering techniques including feature selection and augmentation of transformed features afforded slight improvements to each model's performance.
- A combination of feature creation, selection, and transformation boosted the F1-score of the baseline model by 11%
- Future work could benefit from implementation of MLP progressive imputation, 1D CNNs for binary classifications, and experimenting with feature engineering

References

1. Majidi M, Khadembashi N, Etemad K, et al (2019) Associated factors with major depression: a path analysis on NHANES 2013–2014 study. Int J Cult Ment Health. https://doi.org/10.1080/17542863.2018.1563623
2. Liu M, Wu L, Yao S (2016) Dose-response association of screen time-based sedentary behaviour in children and adolescents and depression: a meta-analysis of observational studies. Br J Sports Med 50:1252–1258. https://doi.org/10.1136/bjsports-2015-095084
3. Liu Y, Ozodiegwu ID, Yu Y, et al (2017) An association of health behaviors with depression and metabolic risks: Data from 2007 to 2014 U.S. National Health and Nutrition Examination Survey. J Affect Disord 217:190–196. https://doi.org/10.1016/j.jad.2017.04.009
4. Patel PO, Patel MR, Baptist AP (2017) Depression and Asthma Outcomes in Older Adults: Results from the National Health and Nutrition Examination Survey. J Allergy Clin Immunol Pract 5:1691-1697.e1. https://doi.org/10.1016/j.jaip.2017.03.034
5. Iranpour S, Sabour S (2019) Inverse association between caffeine intake and depressive symptoms in US adults: data from National Health and Nutrition Examination Survey (NHANES) 2005–2006. Psychiatry Res 271:732–739. https://doi.org/10.1016/j.psychres.2018.11.004
6. Leung CW, Epel ES, Willett WC, et al (2015) Household Food Insecurity Is Positively Associated with Depression among Low-Income Supplemental Nutrition Assistance Program Participants and Income-Eligible Nonparticipants. J Nutr 145:622–627. https://doi.org/10.3945/jn.114.199414
7. Hvas A-M, Juul S, Bech P, Nexø E (2004) Vitamin B₆ Level Is Associated with Symptoms of Depression. Psychother Psychosom 73:340–343. https://doi.org/10.1159/000080386
8. Jerez JM, Molina I, García-Laencina PJ, et al (2010) Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med 50:105–115. https://doi.org/10.1016/j.artmed.2010.05.002
9. Heaton J (2016) An empirical analysis of feature engineering for predictive modeling. In: Conference Proceedings - IEEE SOUTHEASTCON. Institute of Electrical and Electronics Engineers Inc.
10. Yan H, Jiang Y, Zheng J, et al (2006) A multilayer perceptron-based medical decision support system for heart disease diagnosis. Expert Syst Appl 30:272–281. https://doi.org/10.1016/j.eswa.2005.07.022
11. Zanella-Calzada L, Galván-Tejada C, Chávez-Lamas N, et al (2018) Deep Artificial Neural Networks for the Diagnostic of Caries Using Socioeconomic and Nutritional Features as Determinants: Data from NHANES 2013–2014. Bioengineering 5:47. https://doi.org/10.3390/bioengineering5020047
12. Chiu J-S, Li Y-C, Yu F-C, Wang Y-F (2006) Applying an Artificial Neural Network to Predict Osteoporosis in the Elderly
13. Lemineur G, Harba R, Kilic N, et al (2007) Efficient estimation of osteoporosis using Artificial Neural Networks. In: IECON Proceedings (Industrial Electronics Conference). pp 3039–3044
14. Dutta A, Batabyal T, Basu M, Acton ST An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction 15. SVM diabetes
16. Khalilia M, Chakraborty S, Popescu M (2011) Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 11:51. https://doi.org/10.1186/1472-6947-11-51
17. Wyatt MR, Papas M, Johnston T, Taufer M (2016) Development of a scalable method for creating food groups using the NHANES dataset and MapReduce. In: ACM-BCB 2016 - 7th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. Association for Computing Machinery, Inc, New York, NY, USA, pp 118–127

Code available at: https://github.com/csklaver/Capstone-Group6

Email: csklaver@gwu.edu