Project Document: Deep Learning
DeepThoughts

Achilles Rasquinha
David Gao
Chelsea Wu

June $26^{th}, 2021$

Contents

Abstract

Recent advances in the field of Deep Learning have helped us improve using modelling techniques when dealing with extremely raw and highly unstructured data. We aim to utilize such methods with a significant difference in the structures of its input data and attempt to utilize modern deep learning architectures in order to improve the overall prediction. This study explored deep learning methods to predict site-specific SENSE-profitability using 2015 and 2016 SENSE project data (SENSE refers to Sensors for Efficient Nitrogen Use and Stewardship of Environment). Data processing on SENSE experiment data resulted in a large data matrix with over 12000 rows and 100+ columns (100+ features and one target output). We tested various data prepossessing strategies and a few neural network architectures.We observed in general that an increase in number of hidden layers doesn't necessarily increase the overall accuracy. Batch Normalization works best when the batch size is greater or equal to 32 and finally, dropouts on networks where the number of parameters are almost proportional to the number of data points diminishes in performance overall. A model, with a KNN Imputation Strategy, and a layer configuration of 2 x 100 with Batch Normalization Layers outperformed the rest with a good testing accuracy of 82.56%.

Chapter 1

Milestone 1: Project Ideas

1.1 Introduction

We are going to describe our two project proposals – Predicting a site-specific SENSE profitability and Acoustic identification of various “bird calls” in soundscape recordings. Project Idea 1 comes from a currently ongoing research problem in precision agriculture whereas Project Idea 2 comes from a publicly announced prediction challenge and an ongoing research in the field of ornithology. We will attempt to describe and analyze these two project ideas, state a few currently available approaches and estimate the cumulative project influence as well as its potential applications.

1.2 Project Idea 1: Predicting a site-specific SENSE-profitability

SENSE-profitability defines if a site is capable of achieving better economic return for implementing sensor-based nitrogen application practices than implementing conventional grower nitrogen application practices in crop management.

Preliminary investigations found that the effectiveness of sensor-based nitrogen management varied from a site to another. For example, a sensor-based approach may work better on the sites with large variance compared to those relatively uninformed sites. A thorough understanding of relationships between site conditions and the effectiveness of the sensor-based management is critical in determining whether and how to promote the adoption of sensor-based nitrogen management.

In this study, we would like to explore the potential of using deep learning methods to predict whether it is profitable to use sensor-based N application method for a given site with known conventional N application practices and site conditions. We will use research data from project SENSE conducted in

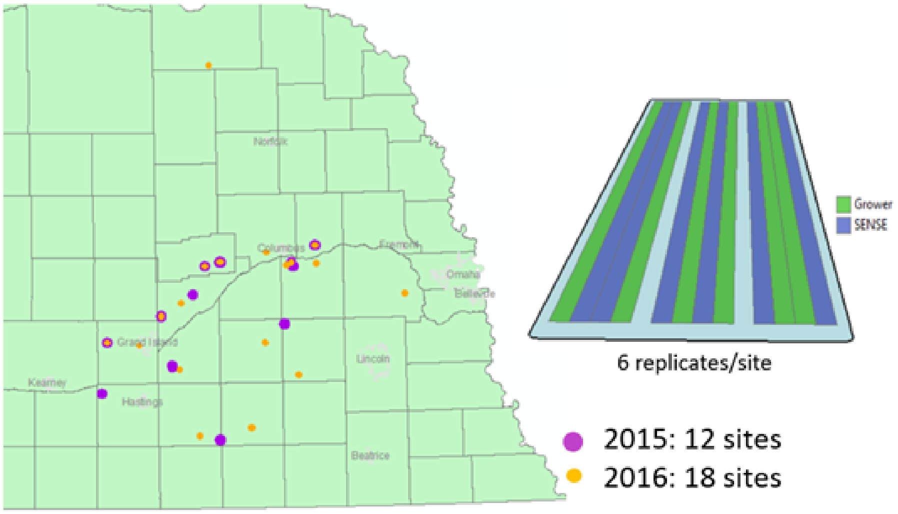


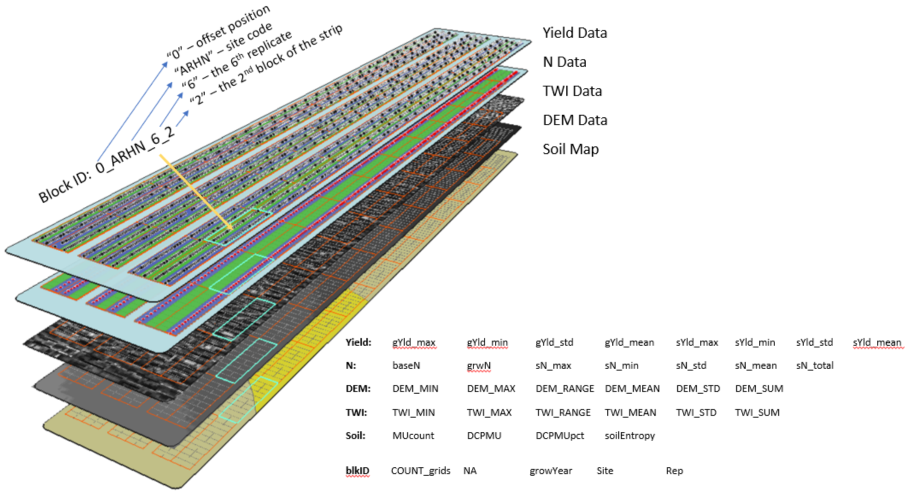
Figure 1.1: 2015 and 2016 SENSE study sites (left) and a sample plot layout (right).

2015 and 2016. Figure 3.1 shows the locations of study sites and a sample plot layout. Features characterizing site conditions were constructed from Digital Elevation Model (DEM) data [9] and soil data. Soil data was from the SSURGO database [8]. Major data components are shown in Figure 1.2.

1.3 Project Idea 2: Acoustic identification of various "bird calls" in soundscape recordings

1.3.1 Introduction

According to a recent study led by the American Museum of National History [2], there are over 18,000+ bird species around the world. Identification of bird species helps to encourage observing numerous patterns in global diversity, their evolutionary history as well as its preservation. They also play a significant role in preserving the overall ecology of a region. For example, guano – the dung accumulated by the excrement of seabirds acts as a natural fertilizer for the soil, thereby acting as a key growth media for plants in that area. The field of ornithology (the methodological study and identification of bird species) involves birdwatchers worldwide aiming to detect various bird species using their ears more than their eyes. Hence, such a study includes a major auditory component when it comes to distinguishing different bird calls. With recent advancements in the field of Deep Learning and Signal Processing, classifying various sounds



Yield:	gYld_max	gYld_min	gYld_std	gYld_mean	sYld_max	sYld_min	sYld_std	sYld_mean
N:	baseN	grwN	sN_max	sN_min	sN_std	sN_mean	sN_total	
DEM:	DEM_MIN	DEM_MAX	DEM_RANGE	DEM_MEAN	DEM_STD	DEM_SUM		
TWI:	TWI_MIN	TWI_MAX	TWI_RANGE	TWI_MEAN	TWI_STD	TWI_SUM		
Soil:	MUcount	DCPMU	DCPMUpct	soilEntropy				
blkID	COUNT_grids	NA	growYear	Site	Rep			

Figure 1.2: Data aggregated to study block.

has now become a reality and is seen in applications such as speech recognition, music classification, speech-to-text, etc. Therefore, we opt to use such state-of-the-art techniques in automating the acoustic identification of bird calls in soundscape recordings effectively, thereby helping ornithologists worldwide to continue maintaining efforts for conservation of biodiversity in birds.

1.3.2 Problem Statement

Given a soundscape recording (as heard within the actual ecological environment perceived by humans, in context), our main objective is to identify the species of birds calling in such recording. The goal is to build an end-to-end application pipeline that considers an input recording unheard of, preprocesses it and detects the various names of bird species the model attempts to identify. Ideally, the input recording would be in an `.ogg` audio format but it could also be of any other universally available and recognized audio format as well. For our feature set, much of the data derived can be digitally downloaded from xeno-canto (XC) – a community driven collaborative project consisting of 6+ million short audio samples of over 10,000+ bird species generously provided and identified by researchers, birders and users all across the globe. XC also claims to have audio recordings of almost over 50% of all identified bird species to date. Also, data provided by eBird for other rich metadata (such as digital photographs, location of such species, etc.) to visually describe more information within our application will be used as well.

1.3.3 Approach

Birdcall Identification problem has been around since the year 2014 to the general public through the LifeCLEF Bird Recognition Challenge. Each year, the competition builds on its previous year by adding new soundscape recordings from different locations alongwith richer metadata in an attempt to challenge the community with a more in depth analysis of the released data and an aim to build better prediction models with a high precision and recall. A survey from many past competitions shows us the usage of recently advanced Convolutional Neural Network (CNN) architectures such as Inception, ResNet or a combination of multiple hybrid architectures for the detection of various birdcalls [4]. A similar public challenge was proposed by the Machine Listening Lab at Queen Mary University of London with challengers offering similar approaches [5]. Much work in the field of birdcall identification has also been pushed forward by the Cornell Lab of Ornithology's Center for Conservation Bioacoustics (CCB) in order to help conservation of wildlife and habitats. This project will attempt to use such state-of-the-art architectures in combination with noise reduction, data augmentation and an in depth feature extraction and representation of audio samples in hope to use additionally rich metadata provided for increased accuracy in detection.

1.3.4 Application

BirdGenie™, is a commercially available mobile application that helps users to identify bird calls on their cellphone devices. Walblr, a UK-based mobile application works on similar lines by helping to identify over 200+ British species. Identification of birdcalls helps researchers to understand changes in ecological systems, overall pollution levels and the possibility of restoration of biodiversity once lost. Conservationists can also install recording units across a region of interest to automate the acoustic detection of birdcalls and thereby improve the conservation efforts and overall quality of life for both birds and humans.

1.4 Conclusions

As mentioned, project idea 1 comes from a currently ongoing research problem in precision agriculture. A previous analysis performed processed the raw data into a well-defined labeled data matrix. If time permits, in addition to implementing deep learning methods to build a model from available data to predict if implementation of sensor-based N application practices is profitable for a site with certain site conditions, we would also like to explore the possibility to build such a model directly from raw data retrieved. Project Idea 2 comes from a research idea currently being explored around the community of ornithologists and data scientists worldwide. The main objective would be an attempt to try improving the accuracy rate of currently available models for such detection using modern available network architectures and signal processing techniques. An end-to-end full-fledged application pipeline would help

Table 1.1: Contributions by team member for Milestone 1.

Team Member	**Contribution**
Chelsea Wu	Project Idea 1
Achilles Rasquinha	Project Idea 2
David Gao	Project Idea 2

research scientists, birdwatchers and the bird-enthusiast community to engage in collaborative efforts for better bird identification and overall conservation.

Chapter 2

Milestone 2: Project Selection

2.1 Introduction

In case of selecting a problem, we intend to choose Project Idea 1: Predicting a site-specific SENSE-profitability. We believe that although the idea is a straight forward classification problem which could be solved with a Basic Feed Forward Neural Network, our goal is to explore the possibility of efficiently optimizing a Deep Learning model using highly-limited source of data, delve into the insights of how robust hyper parameters of our model affects our generalized accuracy and also prove an exhaustive pipeline through our workflow that details from Data Collection methods, to Feature Engineering, to Model Building.

2.2 Problem Specification

Project SENSE is a massive project undertaken by a partnership with the University of Nebraska-Lincoln (UNL), the Nebraska Corn Board, and Nebraska Natural Resources Districts with an intention to improve the nitrogen use efficiency and farmer profitability by adapting the sensor-based nitrogen management practice in farming. SENSE stands for Sensors for Efficient Nitrogen Use and Stewardship of the Environment. SENSE profitability is a term to describe if a farm site is able to achieve better economic return from implementing SENSE nitrogen management practices than Grower's conventional N management. SENSE nitrogen management, in contrast to grower's conventional nitrogen management applies nitrogen uniformly across the field during grow season, uses sensors-based nitrogen application system to detect and evaluate crop's grow condition and then applies an optimal amount of nitrogen where it is needed. The SENSE project team conducted studies in over 30 farm sites to allow growers to test out adopting SENSE N management in their farming.

SENSE N management can improve N use efficiency [7], however, growers will have to update their current N application system to sensor-based system. Knowing up-front how likely using SENSE N management can improve their benefit return is critical for making the decision whether or not to adopt SENSE method in their farming.

2.2.1 Data Source

We will use much of the research data from Project SENSE conducted in 2015 and 2016. Figure 3.1 shows the locations of study sites and a sample plot layout. Each study site has six replicates of comparative N application study: traditional N application practices were conducted in Grower Strips, while sensor-based variable rate N application practices were applied for SENSE strips. Yield data and N application data were collected during project SENSE field experiment. Features characterizing site conditions were constructed from Digital Elevation Model (DEM) data [9] and soil data. Soil data has been obtained from the SSURGO database [8].

1. Study Units

   We partitioned the study strips into 20-feet long grids (width = swath width) to index the study area and associate the N application data and the yield data with the grids. As the yield monitor data is not accurate at the individual point scale, a common practice is to use the value averaged from the yield data collected over a long strip to estimate the field yield rate. Researchers suggest a distance ranged from 70 meters to 90 meters [1][3][6]. We generate 200-feet long blocks based on 20-feet long grids to process the N data, the yield data, and the site condition data. One partitioning generated around 1,300 aggregation blocks. Using a continues integration strategy, shifting aggregation blocks one grid along study strip, we created 9 more sets of aggregation blocks. The overall aggregation blocks then increased from original 1300 to 12000. Summarizing data to aggregation blocks resulted in a data matrix with 75 features and a target output (SENSE-profitability).

2. Features and Target

   The target SENSE profitability is estimated as the profit differences between SENSE and Grower treatments (without counting for the investment for switching from the Grower method to the SENSE method). We use the marginal net return (MNR) to represent profit. MNR can be calculated as the difference between yield income and the N cost (assuming the corn price is 3.83 (\$/bu), and the N price is 0.36 (\$/lb) for both 2015 and 2016):

   $$MNR\ (\$/ac)\ =\ Yield\ (bu/ac)\ *\ 3.83\ (\$/bu)\ -\ N\ (lb/ac) * 0.36\ (\$/ac)$$

   The SENSE profitability then can be calculated as:

   $$diffMNR = sMNR – gMNR$$

where sMNR is the MNR for the SENSE method, and gMNR is the MNR for the Grower method. To explore the prediction problem as classification problem, we create another target feature $diffMNR\_class$ and assign the +1 to those instances with $diffMNR > 0$, whereas the remaining samples with $diffMNR \leq 0$ as -1.

The 75 features are shown in Table 2.1

soilEntropy is a term to characterize how complicated the soil composition within an aggregation block. soilEntropy is calculated as:

$$soilEntropy \; = -\sum_{i=1}^{n} (\frac{A\_{MUi}}{A\_b}) * log(\frac{A\_{MUi}}{A\_b})$$

where A_{MUi} is the total area of soil type MUi, A_b is the study block area, and n is the number of soil types within the block.

The more complex the soil combination is in a study block, the higher its soilEntropy value is. For example, $soilEntropy = 0$ for a study block with only 1 type of soil, $soilEntropy = 0.3$ for a block with two types of soil (50% each), and $soilEntropy = 0.6$ for a block with 4 types of soil (25% each).

2.2.2 Data Exploring

In this section, we discover the feature details for our project SENSE crop data. We use pandas data process library to load the data source file into DataFrame. There are columns which are calculated based on the raw values.

2.2.3 Data Pre-processing

The non-prepossessed data have columns which are highly correlated to target and un-testable data which are generated from script. We have to read the documentation thoroughly as well as the script for generating data source file. Features before prepossessing: Table 2.1 Features after prepossessing: Table 2.2

2.2.4 Data Distribution

Before choosing/applying the methods, we need to get to know our data well first. In this section, we are going to visualize the data distribution towards profitable(+1) or nonprofitable(-1). We will use the data columns after dropping. For the non-numerical data, we will use one-hot encoding to get numerical data with extra columns.

Distribution: Pie Chart

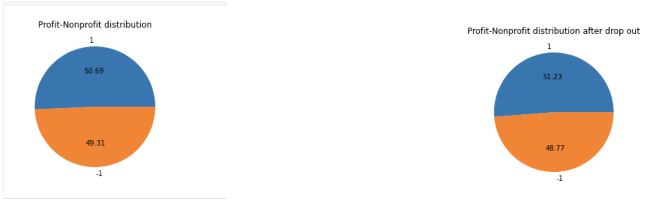
The profit(+1) nonprofit data points numbers from before data cleaning and after are not significant (±1%) 2.1a, figure 2.1b

Table 2.1: Features.

Features	**Description**
gN and baseN	2 features characterizing growers' general N application practices
DEM_MIN, DEM_MAX, DEM_RANGE, DEM_MEAN, DEM_STD, DEM_SUM	statistical features summarized from DEM
TWI_MIN, TWI_MAX, TWI_RANGE, TWI_MEAN, TWI_STD, and TWI_SUM	statistic features summarized from TWI
MU_count, MU_dcp, MU_pct, and soilEntropy	4 features quantifying the complexity of soil combination: MU_count: number of soil types; MU_dcp: dominant soil type; MU_pct: percentage of dominant soil type; soilEntropy
silt, clay, loam, sandy, silty, loamy, very fine, fine, eroded, complex, flooded, slope_l, slope_h, and slope_m	soil descriptive features extracted from soil type description from SSURGO database
soc0_5, soc5_20, soc20_50,soc50_100, soc100_150, soc150_999	Soil Organic Carbon at 6 depths (0 to 5 cm, 5 to 20 cm, 20 to 50 cm, 50 to 100 cm, 100 to 150 cm, and >150 cm)
Silt_DCP_0to5, Silt_DCP_5to20, Silt_DCP_20to50,Silt_DCP_50to100, Silt_DCP_100to150, Silt_DCP_150to999	Percentage of Silt at 6 depths
Sand_DCP_0to5, Sand_DCP_5to20, Sand_DCP_20to50, Sand_DCP_50to100, Sand_DCP_100to150, Sand_DCP_150to999	Percentage of Sand at 6 depths
Clay_DCP_0to5, Clay_DCP_5to20, Clay_DCP_20to50, Clay_DCP_50to100, Clay_DCP_100to150, Clay_DCP_150to999	Percentage of Clay at 6 depths
OM_DCP_0to5, OM_DCP_5to20, OM_DCP_20to50, OM_DCP_50to100, OM_DCP_100to150, OM_DCP_150to999	Organic Matter at 6 depths
AWS_WTA_0to5, AWS_WTA_5to20, AWS_WTA_20to50, AWS_WTA_50to100, AWS_WTA_100to150, AWS_WTA_150to999	Available Water Storage at 6 depths
AWC_DCP_0to5, AWC_DCP_5to20, AWC_DCP_20to50, AWC_DCP_50to100, AWC_DCP_100to150, and AWC_DCP_150to999	Available Water Capacity at 6 depths

Table 2.2: After Drop

Features After dropping	**Correlation(descending)**	**Attribute Description**
baseN	0.246390	Base Nitrogen (N Rate lbs/ac) applied uniformly over the entire site(same for grower and sense strips) before planting season
OM-DCP-150to999	0.181149	Organic Matter (OM)
sandy	0.171202	Soil texture keyword
soc150-999	0.168771	Soil Organic Carbon(g C per square meter) in soil horizon($> 150cm$)
soc100-150	0.166477	Soil Organic Carbon(g C per square meter) in soil horizon($100 \rightarrow 150cm$)
AWC-DCP-150to999	0.139094	Available Water Capacity (AWC) at soil horizon ($150 \rightarrow 999cm$)
DEM-MAX	0.137583	Max value of DEM (elebvation)
OM-DCP-100to150	0.134447	Organic Matter (OM) ($100 \rightarrow 150cm$)
AWS-WTA-100to150	0.132786	Available Water Storage (AWS) ($100 \rightarrow 150$)
AWC-DCP-100to150	0.132564	Available Water Capacity (AWC) at soil horizon ($0 \rightarrow 5cm$)



(a) Before cleaning

(b) After cleaning

Figure 2.1: Profit-nonprofit distribution before and after

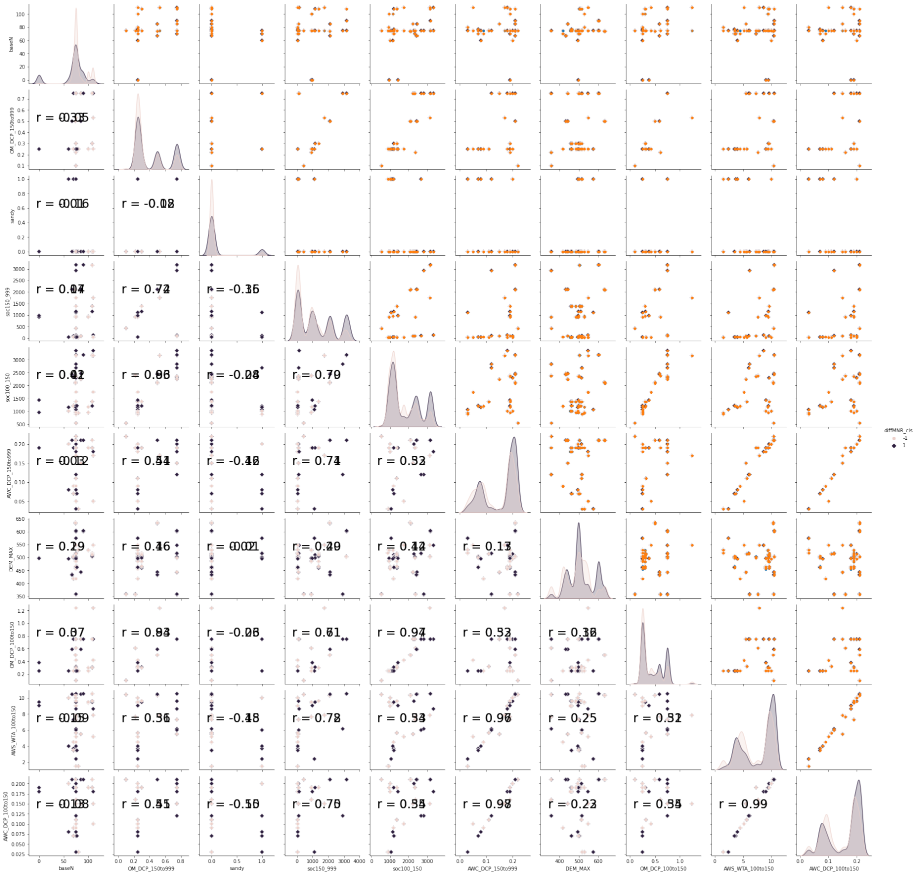


Figure 2.2: Features Scatter Density

Distribution: Mixed Chart

See feature data point distribution in Figure 2.2

Distribution: Individual Chart

See density distribution histogram in Figures 2.3a, 2.3b, 2.3c, 2.3d, 2.3e, 2.3f, 2.3g, 2.3h, 2.3i, 2.3j

2.3 Proposed Method 1: Inverse Correlated Feature

After dropping the unrelated/redundant columns, the correlation between target and other features are not high (way below 0.5), instead of removing the features, we need to take them.

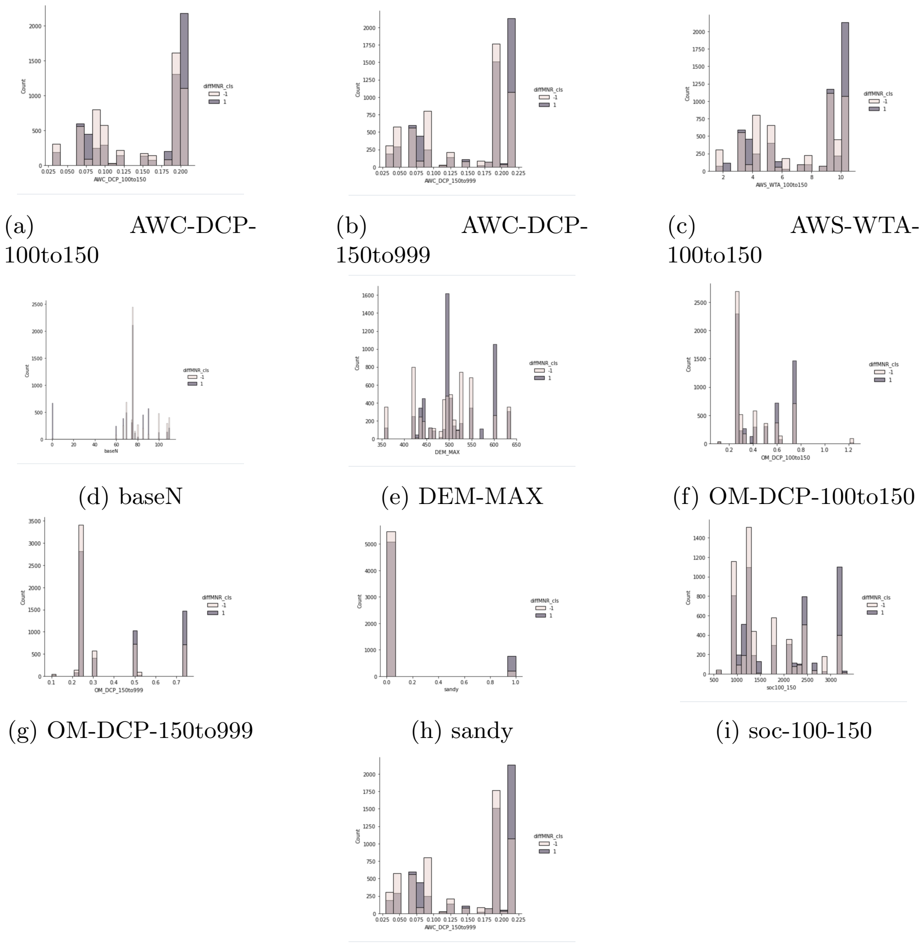
2.4 Proposed Method 2: Multi-Layer Perception Learning

We attempt to provide an end-to-end Deep Learning pipeline that would perform the following tasks:

For data preprocessing, we intend to perform a Data Munging process by ensuring that we estimate the values in places where data is missing. Since the number of samples associated to our given problem is limited, we aim to use the following strategies in order to perform a data imputation task - Simple Imputation (by replacing the mean or median), KNN Imputation (using the k-Nearest Neighbour Clustering algorithm to attempt to fill a missing value), and other imputation strategies provided by the library - fancyimpute. Next, we aim to also perform an outlier detection in order to drop various outliers primarily using a supervised learning algorithm such as a One Class Support Vector Machine approach and other novelty detection algorithms. Finally, we attempt to scale our data using a standardized scaler using linear data transformations or a non-linear scaler. Considering multiple possibilities of many data preprocessing techniques, we aim to see what combination of techniques fit best and provide detailed accuracy results on what choices of prepossessing techniques leads to a better minimized error.

Since the number of samples within our data set are imbalanced on the slight end, we will be attempting to use a Stratified K-Fold Cross-Validation approach that considers K samples as training data and the rest as a test data set, ensuring that the sample distribution within the K samples continue to represent the distribution of the overall data set.

Our objective is to also track the validation loss in real-time and see at what rate does the model attempt to learn starting from a small ratio of training set and building our way towards a larger K over time. We will also tune the



(a) AWC-DCP-100to150

(b) AWC-DCP-150to999

(c) AWS-WTA-100to150

(d) baseN

(e) DEM-MAX

(f) OM-DCP-100to150

(g) OM-DCP-150to999

(h) sandy

(i) soc-100-150

(j) soc-150-999

Figure 2.3: Correlated Columns Distribution

Table 2.3: Timeline.

Task	**Deadline**
Data Preprocessing (Data Munging, Imputation, Outlier Detection and Scaling)	June 15th
Feature Selection (PC-Analysis)	June 19th
Building the Model Framework Pipeline and Architecture	June 21
Evaluating Model Accuracy and Building Classification Report	June 24th

following hyper parameters associated to a given network - the learning rate, the kind of optimizer suitable for learning, the batch sizes associated to each fold, the number of epochs and the number of neurons in each hidden layer along with the depth of the network as well. Our objective is to ensure that the model fits well to the data set provided and hence, we will also attempt to inculcate dropout layers to regularize and avoid over fitting the overall data. We also aim to build multiple model architectures and use an ensemble of many learning architectures to provide an aggregate prediction.

Our intented model will be evaluated based on minimizing the cross categorical entropy loss. Since the estimated prediction is based upon the +1 or -1 value, the objective will be to evaluate the precision, accuracy, F1-score and support values for each of the predicted classes. We also intend to provide the AUC (Area Under Curve) score for the model built.

Since we will be using merely the computational resources provided by HCC (Holland Computing Centre) and is available at free for students and researchers, the cost associated to this task would not be high. However, the timeline associated for the same would be as follows:

2.5 Proposed Method 3: Data Engineering

Current data uses four features to characterize soil composition in an aggregation block: number of soil type (*MU_count*), dominant soil type (*MU_dcp*), percentage of dominant soil type (*MU_pct*), and soilEntropy. Information was lost during the aggregation for those blocks with two or more types of soil. we are thinking to expand the features to include the second type of soil. Then the soil composition features will include: *MU_count*, *MU_dcp*_1, *MU_pct*_1,*MU_dcp*_2, *MU_pct*_2, and soilEntropy.

The soil features (the 57 features including 15 descriptive features and 42 quantitative features) were previously associated to *MU_dcp*. We can experimenting associate each 55 features to both *MU_dcp*_1 and *MU_dcp*_2, which will increase the soil features from current 55 features to 110, or aggregate the two set of features to one set of 55 features.

Current version of data uses 200-feet long aggregation blocks. We can generate more data with various aggregation block sizes(for example, 300-ft long blocks, 400-ft long blocks, block at the full length of study strips) and test see if it will affect the training and test accuracy.

2.6 Conclusions

For this class project, we decide to proceed with the project idea 1 to predict SENSE-profitability using project SENSE study data. The challenge of this project lays in the fact that data was from limited number of farm sites. How to limit over-fitting with limited training samples becomes the main focus of this project. We would like to try the following approaches to limit over-fitting and improve test accuracy:

1. Conduct feature engineering to select a subset of features or extract more features (e.g., expand soil composition features to include the second type of soil in an aggregation block) and test how the feature reduction and expansion affect the modeling.

2. Test out some data pre-processing methods e.g. imputing missing values, detecting and removing outliers, and data regularization etc.

3. Experiment some earlier stop training and randomly dropout methods. Since the data are generating by shifting partial blocks and the difference between the adjacent blocks are not significant. There are a lot of data samples which will provide the same information and lead to unusual high training accuracy. To deal with this, we randomly drop some of the data points(dropout).

4. Build multiple model architectures and use an ensemble of many learning architectures to provide an aggregate prediction.

5. If time permits, we also would like to try autoencoder to create synthetic testing samples to evaluate the model generalization accuracy.

2.7 List of questions for instructor and TA

1. Our group had a meeting with Dr. Scott on June 7th and had all our questions answered.

Table 2.4: Contributions by team member for Milestone 2.

Team Member	**Contribution**
Chelsea Wu	Problem Specification; Data Source;Propose Method 3; Conclusion
David Gao, Achilles Rasquinha	Introduction, Proposed Method 1& 2, Data Exploring, Data Pre-processing, Data Distribution

Chapter 3

Milestone 3: Progress Report

3.1 Introduction

The objective of the study is to predict the "SENSE-profitability" (the capability of achieving better economic return for using sensor-based nitrogen application practices) for a given farm site whose conditions were characterised by soil properties and conventional grower N (nitrogen) application practices. The idea and data derived is based on a research project at the University of Nebraska-Lincoln named Project SENSE. We aim to use a traditional Neural Network model that has helped us achieve an accuracy of 70%.

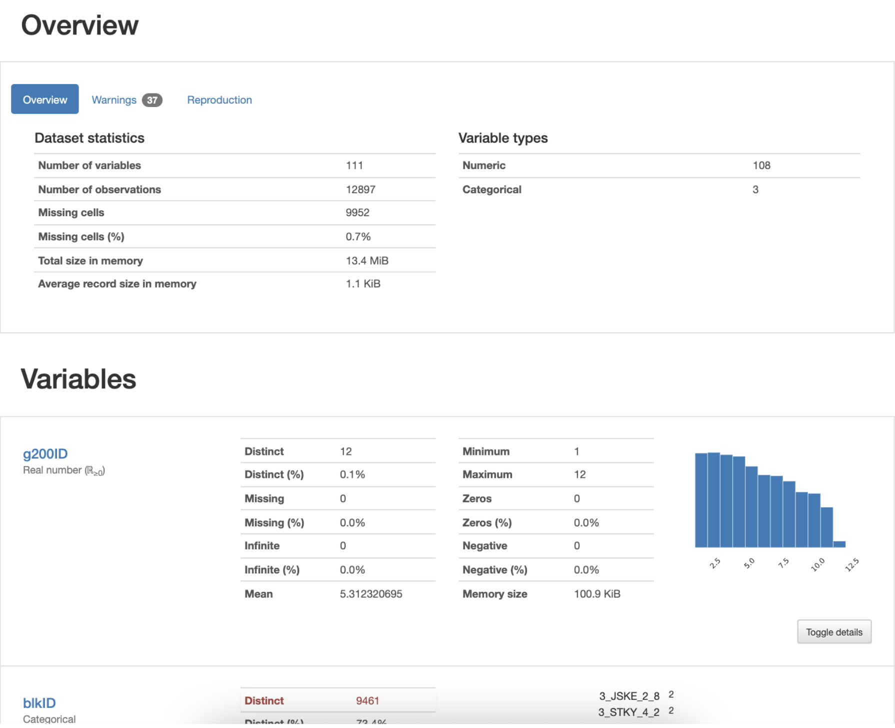
3.2 Experimental Setup

3.2.1 Data Sources

We use the data processing library "pandas" to load raw data into a pandas DataFrame object. Our data is currently available on this link which contains information of all soil data since the year 2015 to 2016 in sites positioned within Nebraska.

3.2.2 Data Profile

We use "pandas-profiling" for profiling our data. We receive an impressive amount of valuable information regarding the nature of the data starting with the number of missing data, the statistical distribution of such information and its impact on how data could be further processed. An example of the report specifications generated can be viewed in Figure 3.1



Overview

Dataset statistics

Number of variables	111
Number of observations	12897
Missing cells	9952
Missing cells (%)	0.7%
Total size in memory	13.4 MiB
Average record size in memory	1.1 KiB

Variable types

Numeric	108
Categorical	3

Variables

g200ID
Real number ($\mathbb{R}_{\geq 0}$)

Distinct	12	Minimum	1
Distinct (%)	0.1%	Maximum	12
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	5.312320695	Memory size	100.9 KiB

blkID
Categorical

Distinct	9461
Distinct (%)	73.4%

3_JSKE_2_8	2
3_STKY_4_2	2

Figure 3.1: Data Profile summary

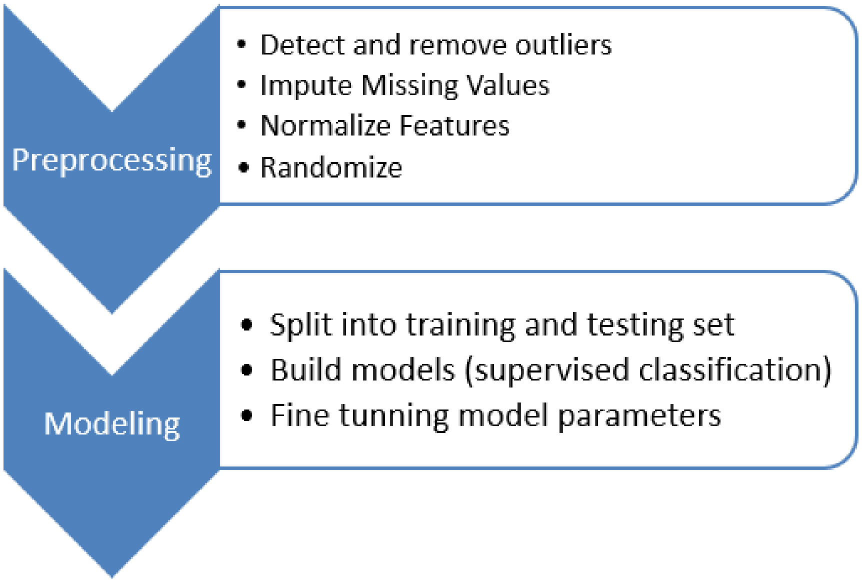

Preprocessing
• Detect and remove outliers
• Impute Missing Values
• Normalize Features
• Randomize
Modeling
• Split into training and testing set
• Build models (supervised classification)
• Fine tunning model parameters

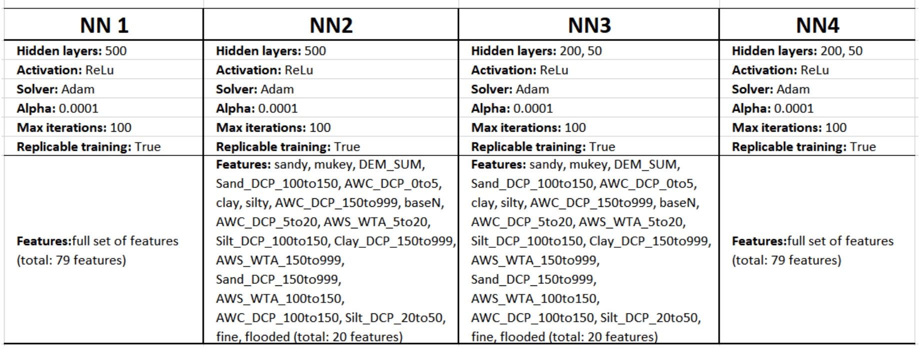
Figure 3.2: Data Modeling Pipeline

3.2.3 Data Prepossessing

We plug the data (described in milestone 2) into modeling pipeline shown in Figure 3.2 to model the relationship between inputs to output. The data preprocessing first detected and removed outliers (The Isolation Forest method was used with a contamination rate of 1%), then imputed missing values by dropping the rows with missing values. We also standardized all features to $\mu = 0$, $\sigma^2 = 1$. After that, all rows were shuffled to break the existing connections between instance and class. After the preprocess and outlier removal analysis steps, the 12,000+ samples were reduced to 11,402 instances, which then were split into the training data and test data: 70% randomly selected samples (7,982) as the training data, while the rest samples (3,420) were used as the test data.

3.2.4 Architectures and hyper parameter

We tested 4 neural network models NN1, NN2, NN3 and NN4 shown in Figure 3.3. Models NN1 and NN 4 used full set of features (a total of 79 features), whereas models NN2 and NN3 use top 20 most important features.



NN 1	NN2	NN3	NN4
Hidden layers: 500	**Hidden layers:** 500	**Hidden layers:** 200, 50	**Hidden layers:** 200, 50
Activation: ReLu	**Activation:** ReLu	**Activation:** ReLu	**Activation:** ReLu
Solver: Adam	**Solver:** Adam	**Solver:** Adam	**Solver:** Adam
Alpha: 0.0001	**Alpha:** 0.0001	**Alpha:** 0.0001	**Alpha:** 0.0001
Max iterations: 100	**Max iterations:** 100	**Max iterations:** 100	**Max iterations:** 100
Replicable training: True	**Replicable training:** True	**Replicable training:** True	**Replicable training:** True
Features:full set of features (total: 79 features)	**Features:** sandy, mukey, DEM_SUM, Sand_DCP_100to150, AWC_DCP_0to5, clay, silty, AWC_DCP_150to999, baseN, AWC_DCP_5to20, AWS_WTA_5to20, Silt_DCP_100to150, Clay_DCP_150to999, AWS_WTA_150to999, Sand_DCP_150to999, AWS_WTA_100to150, AWC_DCP_100to150, Silt_DCP_20to50, fine, flooded (total: 20 features)	**Features:** sandy, mukey, DEM_SUM, Sand_DCP_100to150, AWC_DCP_0to5, clay, silty, AWC_DCP_150to999, baseN, AWC_DCP_5to20, AWS_WTA_5to20, Silt_DCP_100to150, Clay_DCP_150to999, AWS_WTA_150to999, Sand_DCP_150to999, AWS_WTA_100to150, AWC_DCP_100to150, Silt_DCP_20to50, fine, flooded (total: 20 features)	**Features:**full set of features (total: 79 features)

Figure 3.3: NN Models

Neural Networks

For the Artificial Neural Network model (using a multi-layer perceptron), we cannot create CNN since the internal layer dimension mismatch when the data goes through the Conv-layer. Even our feature data is a multidimensional array, it will require different techniques to deal with from the image data. We decide to use normal neural network. We set the model max iterations to 500 uniformly.

```
max_iter = 500
model_ann = MLPClassifier(activation='relu',
                          solver='adam',
                          max_iter=max_iter)
```

3.2.5 Performance Measures

As displayed within the results mentioned in 3.3.3, we display the training accuracy and the corresponding cross-validation score for each of the learning models we have used. A suitable division of 60% training v/s testing set was used in order to generate a responsive result.

3.2.6 Other techniques to improving performance

Data Prepossessing has attempted to significantly help us improve our overall performance score providing us decent and accurate results for the same.

3.3 Experimental Results

The experimental results from the 4 test models are shown in Table 3.1. As we can see from Table 3.1, the models with full set of features have better performance than the two models with only selected 20 features. We will test

Table 3.1: NN Test Results

	AUC	**CA**	**F1**	**Precision**	**Recall**
NN1	0.87	0.79	0.79	0.79	0.79
NN2	0.79	0.73	0.73	0.73	0.73
NN3	0.80	0.73	0.73	0.73	0.73
NN4	0.89	0.81	0.81	0.81	0.81

more models see if we can find better architectures and better performed models.

Current data uses four features to characterize soil composition in an aggregation block: number of soil type (MU_count), dominant soil type (MU_dcp), percentage of dominant soil type (MU_pct), and soilEntropy. Information was lost during the aggregation for those blocks with two or more types of soil. We expanded the soil properties features in each aggregation block to include the secondary soil type. Over the 12000+ sampled blocks, 2679 blocks contain 2 types of soil, whereas 141 blocks contain 3 types of soil type. Such expansion added MU_dcp_2, MU_pct_2, and a total of 57 features (15 descriptive features and 42 quantitative features) associated with MU_dcp_2. We will try to fit the model using the expanded dataset see whether the expansion of the feature space will improve the accuracy.

3.4 Discussion

Our objective as of now has helped us notice that given a certain amount of data cleansing, preprocessing and scaling - we've achieved a fair amount of accuracy extended. We have also attempted to use other Deep Learning architectures in order to extend our overall accuracy score. This, however, wasn't suited at best due to limitations within the structure of our data for which Convolutions Neural Network architectures didn't seem feasible to an extent. Our objective further would be also to perform an "ensemble learning approach" which is hope to maximize the accuracy using multiple architecture parameters such as the number of hidden layers, and the depth of the layer architecture as well.

3.5 Conclusion

Our source code is currently avaiable on the following links - Notebook 1 and Notebook 2 in order to reproduce our experiment performance and results. One can reproduce our resultant analysis on the said links by simply running the notebook on the destined server. All information regarding data, data profiling reports as well as figures generated are currently available on the links provided.

Table 3.2: Contributions by team member for Milestone 3.

Team Member	**Contribution**
David Gao	Intro,Experimental Setup, Model Result
Achilles Rasquinha	Data Profiling, Model Generation, Discussion and Conclusion
Chelsea Wu	Abstract, processing pipeline, NN models, test results, expand soil features Figure 3.2, 3.3, Table 3.1

Chapter 4

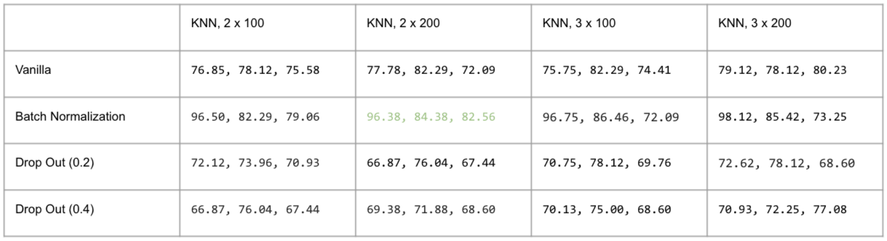
Milestone 4: Final Report

4.1 Introduction

Nitrogen use efficiency has been a long-existing research topic in agriculture, as agricultural nitrogen (N) application is a crucial factor affecting crop yield, and is a common source contaminating groundwater. Sensor-based nitrogen management uses crop canopy sensors to capture N status of living crop in the field then applies N accordingly. Sensor-based N management is expected to have better N Use Efficiency (NUE) than conventional grower's fixed rate N application management. Project SENSE (Sensors for Efficient Nitrogen Use and Stewardship of the Environment) is a project led by the On-Farm Research team of the University of Nebraska-Lincoln (UNL) to improve the NUE and farmer profitability by adapting the sensor-based N management practice in farming. During 2015 and 2016, the SENSE project team conducted studies in over 30 farm sites to allow farmers testing out SENSE-based N management and to compare the differences of the two methods Figure 3.1. SENSE-profitability defines if a site is capable of achieving better economic return for implementing sensor-based method than implementing conventional grower's N management. The purpose of this study is to use deep learning methods to predict site-specific SENSE-profitability: for a given site with known conventional N application practices and site conditions, whether it is profitable to use sensor-based N application method. Knowing up-front how likely using SENSE N management can improve their benefit return is critical for making the decision whether or not to adopt SENSE method in their farming.

This study uses 2015 and 2016 SENSE project data, DEM data from USGS, and soil data from the Natural Resources Conservation Service SSURGO database.

The study strips were partitioned into 20-feet long grids to index the study area and to associate the N application data and the yield data with the grids. 200-feet long blocks were generated based on 20-feet long grids to process the N data, the yield data, and the site condition data. One partitioning generated around 1,300 aggregation blocks. Using a continues integration strategy,



	KNN, 2 x 100	KNN, 2 x 200	KNN, 3 x 100	KNN, 3 x 200
Vanilla	76.85, 78.12, 75.58	77.78, 82.29, 72.09	75.75, 82.29, 74.41	79.12, 78.12, 80.23
Batch Normalization	96.50, 82.29, 79.06	96.38, 84.38, 82.56	96.75, 86.46, 72.09	98.12, 85.42, 73.25
Drop Out (0.2)	72.12, 73.96, 70.93	66.87, 76.04, 67.44	70.75, 78.12, 69.76	72.62, 78.12, 68.60
Drop Out (0.4)	66.87, 76.04, 67.44	69.38, 71.88, 68.60	70.13, 75.00, 68.60	70.93, 72.25, 77.08

Figure 4.1: Evaluation across various hyper parameters.

shifting aggregation blocks one grid along study strip, we created 9 more sets of aggregation blocks. The overall aggregation blocks then increased from original 1300 to 12000. Summarizing data to aggregation blocks resulted in a data matrix with over 100 features (73 features were selected to be used in the modeling) and a target output (SENSE-profitability). The 73 features are shown in Table 2.1. The target SENSE profitability is estimated as the profit differences between SENSE and Grower treatments (without counting for the investment for switching from the Grower method to the SENSE method).

4.2 Experimental Setup

Our model can be reproduced on the following link using Google Colaboratory. As mentioned, the data source used comes a list of 73 features used in Table 2.1. We initially performed an imputation strategy of k-Nearest Neighbors (where k = 5) that was then used to replace all possible missing values within our dataset, since our dataset consisted of limited number of data points. For preprocessing, we used a Minimum-Maximum Scaler between 0 and 1. We also encode our dataset using a One-Hot encoding scheme of features that are ”class like”.

For our model, we ensure that the model evaluates on a dataset split of 60% training data, 20% of validation data and 20% of testing data. In case of our model architecture, we evaluate our entire dataset split with a combination of N x D architecture where N is within the range [2, 3] and D is within the range [100, 200]. We attempt to use Batch Normalization Layers between each hidden layer and the Activation Layer also a Dropout Layer right after an Activation Layer.

4.3 Experimental Results

On evaluating a cross-combination of the above hyper-parameters, we achieve the results mentioned in Figure 4.1

We notice that a 2 x 100 layered architecture with Batch Normalization done in batch sizes of 32 for a range of 50 epochs, a learning rate of 0.001 using an Adam Optimizer and a Binary Cross Entropy Loss function helps us achieve an

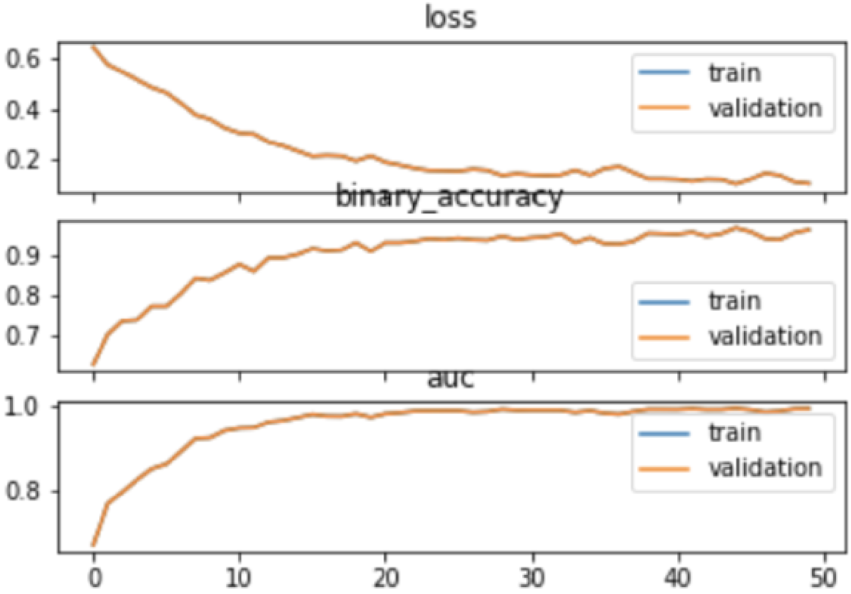


Figure 4.2: Loss, Accuray and AUC graph

impressive training accuracy of 96.38%, a validation accuracy of 84.38% and a testing accuracy of 82.56%. Our network can be visualized in Figure 4.5

Figure 4.2 represents the model performance over time for our defined parameters mentioned above. We notice that the model tends to generalize well over 25+ epochs. There is also a gradual decrease in loss over time as well.

Our generated Confusion Matrix as provided in Figure 4.3 provides us a total precision for cases where Sensor-Based Nitrogen isn't required however performs moderately well (if not best), in case of sites where it shows profitable results. This is more evidently evaluated by the classification report generated in Figure 4.4.

4.4 Discussion

On tweaking much of the hyper parameters, we observed in general that an increase in number of hidden layers doesn't necessarily increase the overall accuracy. Batch Normalization works best when the batch size is greater or equal to 32 and finally, dropouts on networks where the number of parameters are almost proportional to the number of data points diminishes in performance overall.

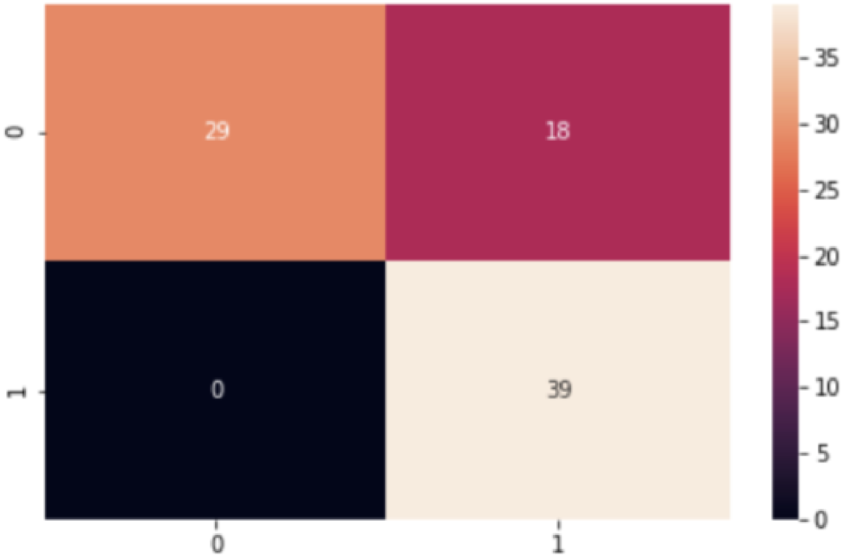
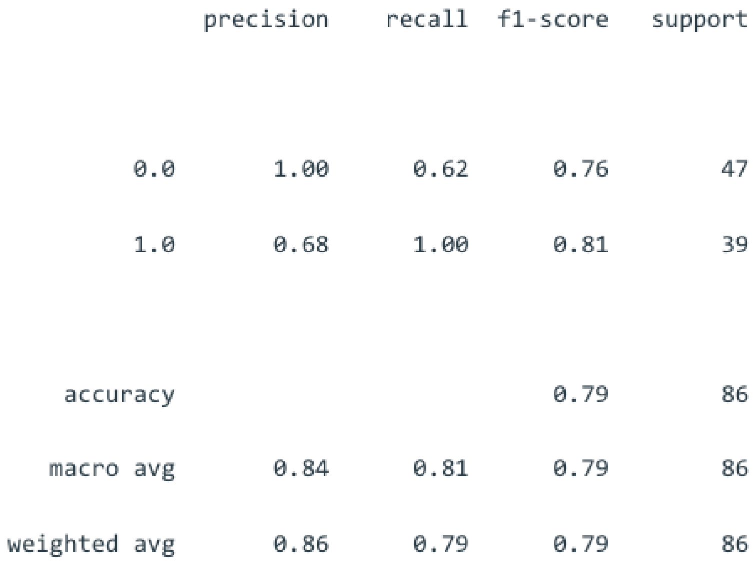


Figure 4.3: Confusion Matrix

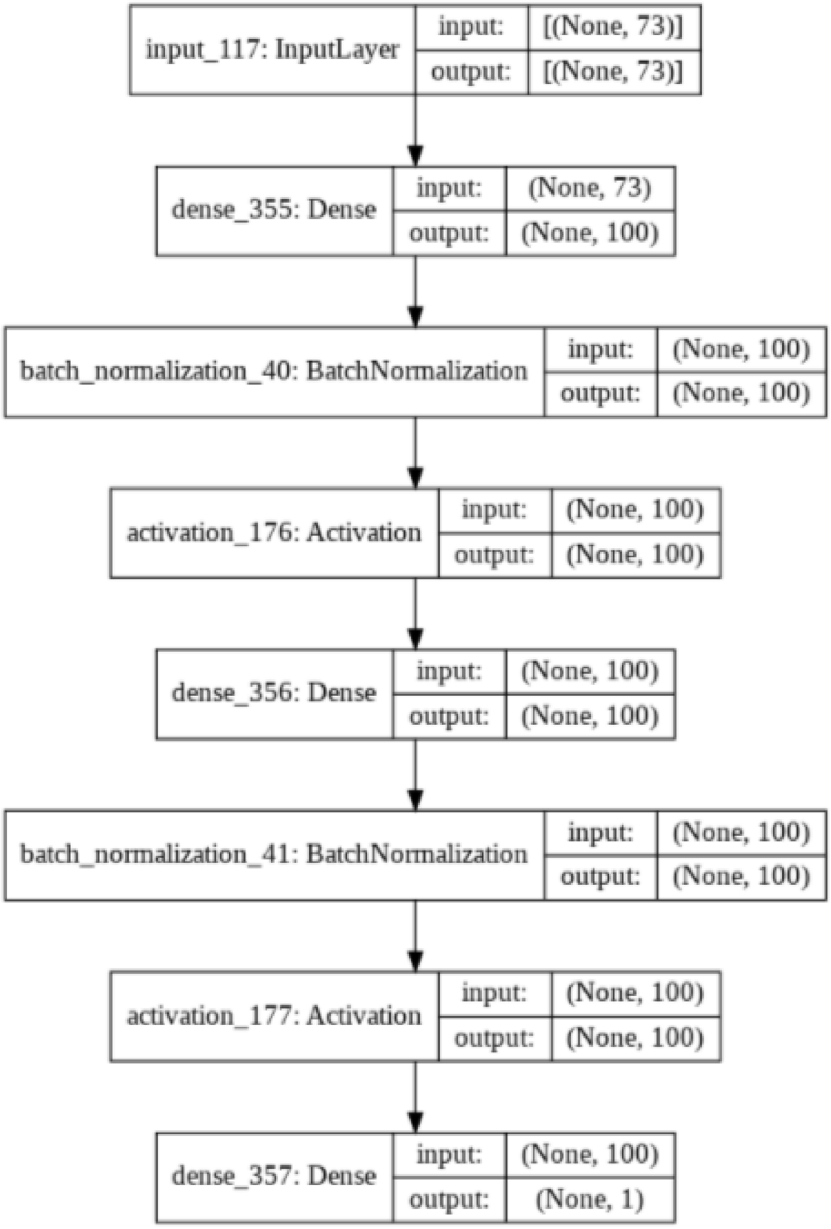


```
              precision    recall  f1-score   support

         0.0       1.00      0.62      0.76        47
         1.0       0.68      1.00      0.81        39

    accuracy                           0.79        86
   macro avg       0.84      0.81      0.79        86
weighted avg       0.86      0.79      0.79        86
```

Figure 4.4: Classification Report



rm

Figure 4.5: Network Graph

Table 4.1: Contributions by team member for Milestone 5.

Team Member	**Contribution**
Achilles Rasquinha	Experimental Setup and Results
Chelsea Wu	Abstract, Introduction, Conclusion
David Gao	Results

4.5 Conclusion

This study explored deep learning methods to predict site-specific SENSE profitability using 2015 and 2016 SENSE project data. The 2015 and 2016 studies' strips were partitioned into small index grids and then were regrouped to 200 feet long study blocks. Site conditions at each study block were characterized with DEM, TWI and soil properties data and were associated to the SENSE-profitability derived from project SENSE study data. Dataset includes over 12,000 samples with over 75 features and one target output.

Our Model, with a KNN Imputation Strategy, and a layer configuration of 2 x 100 with Batch Normalization Layers performed a good testing accuracy of 82.56%.

Due to the limited number of study sites (30 sites), and the way the data was organized, over fitting is most likely unavoidable in this study. In the future, we will test more structures see if we can achieve better accuracy. We will also try Convolution network to take advantage of spatial data in this study and see if it can give better result.

Bibliography

[1] MA Al-Mahasneh and TS Colvin. Verification of yield monitor performance for on-the-go measurement of yield with an in-board electronic scale. *Transactions of the ASAE*, 43(4):801, 2000.

[2] George F. Barrowclough, Joel Cracraft, John Klicka, and Robert M. Zink. How Many Kinds of Birds Are There and Why Does It Matter? *PLOS ONE*, 11(11):e0166307, November 2016. URL: `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0166307`, `doi:10.1371/journal.pone.0166307`.

[3] David S Bullock, Maria Boerngen, Haiying Tao, Bruce Maxwell, Joe D Luck, Luciano Shiratsuchi, Laila Puntel, and Nicolas F Martin. The data-intensive farm management project: Changing agronomic research through on-farm precision experimentation. *Agronomy Journal*, 111(6):2736–2746, 2019.

[4] Stefan Kahl, Mary Clapp, W. Hopping, Hervé Goëau, Hervé Glotin, Robert Planqué, Willem-Pier Vellinga, and Alexis Joly. Overview of birdclef 2020: Bird sound recognition in complex acoustic environments. 09 2020.

[5] Stefan Kahl, Mary Clapp, W. Hopping, Hervé Goëau, Hervé Glotin, Robert Planqué, Willem-Pier Vellinga, and Alexis Joly. Overview of birdclef 2020: Bird sound recognition in complex acoustic environments. 09 2020.

[6] Fernando Perez-Munoz and TS Colvin. Continuous grain yield monitoring. *Transactions of the ASAE*, 39(3):775–783, 1996.

[7] Peter C Scharf, D Kent Shannon, Harlan L Palm, Kenneth A Sudduth, Scott T Drummond, Newell R Kitchen, Larry J Mueller, Victoria C Hubbard, and Luciane F Oliveira. Sensor-based nitrogen applications outperformed producer-chosen rates for corn in on-farm demonstrations. *Agronomy Journal*, 103(6):1683–1691, 2011.

[8] Soil Survey Staff. Gridded soil survey geographic (gssurgo) database for nebraska, 2020. July 20, 2020 (FY2020 official release). URL: `http://datagateway.nrcs.usda.gov/`.

[9] U.S. Geological Survey. 3d elevation program 1-meter resolution digital elevation model, 2020. accessed Jyly 2020. URL: https://viewer.nationalmap.gov/basic/.