C O V E N T R Y
U N I V E R S I T Y

Faculty of Engineering, Environment and Computing
School of Computing, Electronics and Mathematics

MSc. Data Science and Computational Intelligence
7151CEM
Computing Individual Research Project

Analyse the deep learning technique to predict multiple characteristics of the soil without chemical testing for application in soil mapping across different geographies

Author: Student1

SID: 10xxxx

1st Supervisor: Dr. Alireza Daneshkhah
2nd Supervisor: Dr. Tabassom Sedighi

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Data Science & Computational Intelligence

Academic Year: 2020/21

Declaration of Originality

I declare that this project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

Statement of copyright

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialise products and services developed by staff and students. Any revenue that is generated is split with the inventor/s of the product or service. For further information please see www.coventry.ac.uk/ipr or contact ipr@coventry.ac.uk.

Statement of ethical engagement

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking)

Signed: Date: 12 Aug 2021

Please complete all fields.

First Name:	Student 1 first name
Last Name:	Student 1
Student ID number	10xxxxx
Ethics Application Number	P1xxxx
1st Supervisor Name	Dr. Alireza Daneshkhah
2nd Supervisor Name	Dr. Tabassom Sedighi

This form must be completed, scanned, and included with your project submission to Turnitin. Failure to append these declarations may result in your project being rejected for marking.

Abstract

Soil analysis is the science of analysing soil samples and determine various organic and non-organic characteristics of the soil. The agriculture industry has been seeking for innovative non- destructive solutions to regularly monitor the changing physical and chemical properties of the land with a goal to improve agricultural yield, agronomics management and efficient selection of crops based on the soil properties. In addition to this governments are also interested to know the soil characteristics across countries to understand impact of weather and climate and make informed policy decisions. In this context often land surveys are conducted by independent authorities and collected soil samples are lab tested to determine multiple soil properties. Pre-dominantly traditional methods are used for soil analysis but since last decade diffuse reflectance infrared spectroscopy is getting very popular. This technique allows the rapid acquisition of soil information in the field or in the lab however it needs the creation of regional to global soil spectral libraries. Machine learning researchers have also shown interest in applying various techniques on this visual(vis) to near infrared (NIR) spectral data (400nm – 2500nm) region and extract accurate soil information without the need of creation and maintenance of regional to global soil spectral libraries. This research is focussed on the similar goal to study and apply deep neural network methods on raw spectroscopy data and simultaneously predict multiple soil attributes using a single model. I have demonstrated the effectiveness of deep neural network on the LUCAS soil dataset which consists of approximately 21,000 topsoil observations with physical and biological properties from the Europe. I have represented the spectral data in 1D spectrogram as continuous time series to fully utilise the effectiveness of multilayer and CNN models. In addition to using raw spectral data, I have also combined auxiliary information like geographical elevation of sample point, land coverage and land usage. I have showed the capacity of deep neural network methods with Noise removal , Data smoothening and dimensionality reduction of raw spectral data to simultaneously predict 8 soil properties [pH(H_20), pH ($CaCl_2$), EC, OC, N, P, K, $CaCO_3$] using a single model. Root mean squared error (RMSE) of pH(H_20) & pH ($CaCl_2$) are noted to be lowest among all the predicted values as 0.81 and 0.85 respectively. Best RMSE values of EC, OC, P, N,K & $CaCO_3$ are 21.22, 34.46, 27.02 ,1.93 , 142.61 & 23.98 respectively.

Table of Contents

Abstract 2
Table of Contents 3
Acknowledgements 5
1 Introduction 6
1.1 Background to the Project 6
1.2 Project Objectives 7
1.3 Report Structure 7
2 Literature Review 9
3 Methodology 11
4 Requirements 12
4.1 Dataset Selection 12
4.2 Gap analysis 13
4.3 System requirements 13
4.4 Functionality requirements 13
4.5 Non -Functionality requirements 14
5 Dataset 14
6 Design 16
7 Techniques 18
7.1 Feature Scaling 18
7.2 Data Smoothening : Savitzky-Golay 19
7.3 Feature Dimensionality Reduction : Principle Component Analysis (PCA) 19
7.4 Multi-Layer Perceptron (MLP) 19
7.5 Convolutional Neural Network (CNN) 20
8 Implementation 21
8.1 Data Smoothening 21
8.2 Principal Component Analysis (PCA) 22
8.3 MLP Architecture 23
8.4 CNN Architecture 24
8.5 Metrics : Root Mean Squared Error (RMSE) 25
9 Results 26
9.1 Exploratory Data Analysis(EDA) 26
9.2 Deep learning Models (MLP and CNN) 28
10 Project Management 30
10.1 Project Schedule 30

10.2 Risk Management 30
10.3 Quality Management 31
10.4 Social, Legal, Ethical and Professional Considerations 31
11 Conclusions 31
11.1 Achievements 31
11.2 Future Work 32
12 Student Reflections 33
Bibliography and References 34
Appendix A – Project Plan 1
Appendix B – Interim Progress Report and Meeting Records 1
Appendix D – Project Presentation 1
Appendix E – Certificate of Ethics Approval 1

Table of Tables

Table 1 Summary of different research papers 11
Table 2 System Configuration 13
Table 3 Dataset Combinations 13
Table 4 Topsoil Attributes in dataset 14
Table 5 Attributed in Spectral datafiles 15

Table of Figures

Figure 1 : Waterfall model (adopted for this research work) 12
Figure 2: Design building blocks 17
Figure 3: High level design 18
Figure 4 Multi-Layer Perceptron (Abirami et al.,2020) 20
Figure 5 Typical CNN layers (Goodfellow et al., 2021) 21
Figure 6 Before applying Savitzky-Golay filter 21
Figure 7 After applying Savitzky-Golay filter 22
Figure 8 Explained Variance 22
Figure 9 Cumulative explained variance 23
Figure 10 MLP Base Architecture 24
Figure 11 CNN Architectures 25
Figure 12 Outliers in target variable 26
Figure 13 Heatmap of different soil attribute 27
Figure 14 Data distribution of target variables 27
Figure 15 Original Spectral Data 28
Figure 16 Spectral Data After Noise Removal 28
Figure 17 Result RMSE Summary 29

Acknowledgements

I would like to like thanks my supervisor Dr. Alireza Daneshkhah for his full support and guidance during the entire duration of my research. He had been kind enough to provide feedback at every stage of the project and reviewed my progress without any delay. His feedbacks have helped me to improve my work further.

Also, I would like to thank my dear wife for her patience and unconditional support during this entire duration of my studies and keeping me motivated to perform my best.

I also appreciate and thanks every member of the faculty who has taught me in this course and shared their valuable knowledge which has helped me to complete this project.

1 Introduction

Soil ecosystem which is one of the most critical resource on the plant, is getting impacted due to several factors including land development and excessive use of chemical in agriculture. So, it is important than ever to have an efficient and better management of soil which can monitor its physical and chemical properties on regular basis(FAO and ITPS, 2015).
There are many available methods for soil analysis which involve 'wet methods' like chemical testing in laboratory or 'dry methods' which makes use of soil spectral values in the wavelength range of visual to near infrared (vis-NIR)(McBratney et al. 2002).
Due to its benefit of being non-destructive and cost-effective solution, diffuse reflectance spectroscopy, has gained popularity in the past decades.
In this regard, significant soil spectral libraries have been developed by many enthusiasts by carefully cataloguing soil samples and their specific spectral reflectance values and associated properties of the soil. Land Use and Coverage Area Frame Survey (LUCAS) topsoil dataset is one such prominent example in this field (Orgiazzi et al., 2018).
This research is aimed to study the feasibility of deep neural networks to simultaneously predict multiple soil properties from the vis-NIR spectra using a single model. Below sub-sections explain in detail about the project background and its objectives.

1.1 Background to the Project

Land surveys are often conducted by government authorities to capture land coverage, usage, and soil mapping across different areas. One of the important aspect of these land surveys is to assess the soil organic and non-organic properties. Soil samples from different locations are collected using a pre-defined standard method and sent to laboratory for further analysis. The total number of collected soil samples have a direct influence on the survey's results as low number of samples may lead to inaccurate representation of soil mapping whereas high number of sample collection can indirectly add to excess cost and time and hence putting entire campaign at risk and limit number of survey's in future.
Soil properties also play an important role in agriculture as productivity of crops is directly linked with the quality of soil in the field. At a small scale it may seems practical and cost efficient to perform soil testing in laboratory for each sample but as the number of samples increase with the increase in area of land, lab testing becomes costly, and time consuming and hence alternate ways of testing is required for analysing soil properties. In Last decade , use of vis-NIR (visual Near Infrared) spectroscopy has gained popularity (Kopaˇcková & Ben-Dor, 2016) in soil studies because it enables soil surveyors to increase soil sample densities without incurring any additional cost and provide potential solution for soil mapping and monitoring of large areas (Wetterlind et al., 2008;Ramirez-Lopez et al., 2014).
However increased research effort in vis-NIR spectroscopy has resulted in development of many soil databases (i.e., spectral libraries) for the quantification of soil properties at regional (Gogé et al., 2012) or global scales (Viscarra Rossel et al., 2016). The increased number of spectral libraries also bring challenge to maintain those libraries and keep them relevant for future purposes and cost effective.
To address this problem of predicting soil properties using land survey's information and vis-NIR spectral data, deep learning techniques would be explored.

The Insights from land surveys are mostly used by government authorities in formulating the policies around the agriculture, land coverage, land usage or understanding the environmental changes over the period. Government policies towards city development can be influenced based on the information about the soil characteristics. For example, fertile land is more likely to be left for agriculture compared to barren land which is more suited for extending cities and industrial developments.
Soil characteristics information can be helpful to farmers in deciding which type of crop is best suited and will produce maximum yield in a certain region/geography. Also, using the state of

art machine learning algorithms in soil analysis, can reduce overall analysis time and cost significantly. This will increase the number of survey frequency to provide better insight into changes in soil over the time for better management of farmlands.

1.2 Project Objectives

This research project main goal is to assess the feasibility of deep learning techniques in predicting multiple properties of soil from a single model. This study would focus on 2 main research questions as listed below:

1. Can deep learning techniques predict accurate multiple soil characteristics by using vis-
   NIR spectral data?
2. Is vis-NIR spectral data enough to predict soil characteristics from any region/continent or external information related with land coverage and land usage also required for accurate prediction?

1.3 Report Structure

This report is divided into 12 main section with each section focus on a specific set of activities as described below. Additional information and materials are provided in Appendix for reference.

Section 1: This section provides introduction , background of the project and project objectives.

Section 2: This section mainly covers literature review to highlight work of other researchers in this field and helps to understand current progress in this field of research.

Section 3: This section covers project methodology and rationale behind its selection for this project work.

Section 4: This section cover requirement gathering and different approach followed for dataset and requirement finalization.

Section 5: This section provides overview of the input dataset and briefly highlights all its attributes.

Section 6: This section mainly explains the software building blocks and high-level design for this project.

Section 7: This section explains various techniques applied in this research.

Section 8: This section deals with the implementation details for each of the technique applied in section 7.

Section 9: This section present results and discusses finding from EDA and deep learning models.

Section 10: This section covers project management aspects of this research and briefly explains key processes.

Section 11: This section concludes project finding and provide recommendation for future work.

Section 12: This section provides student reflections and covers high and lows of the project.

2 Literature Review

In the past many researchers have carried out studies to find a suitable and appropriate techniques in soil analysis and predominantly Principal component regression (PCR) and Partial least squares regression (PLSR) were found to be used in regression models for soil analysis using spectral wavelength data. This section aims to review some of latest research work involving deep learning techniques and highlight the preferred algorithms and best practices used by researchers in the field of soil analysis. For the comparison, only the papers using the LUCAS dataset and vis-NIR spectral data had been considered for the review.

In 2015, Veres et al. did the research to leverage advances in deep learning architectures mainly Convolutional Neural Network (CNN) & Conditional restricted Boltzmann machines (CRBM) for structured output prediction of soil properties. Authors have used spectrogram data from LUCAS dataset to predict 5 properties of soil namely Sand, pH, Soil Organic Carbon (SOC), Calcium (Ca) & Phosphorous (P). They have considered only spectral measurements as input features for the models and have excluded all the other available auxiliary features information. Since input features were only spectral wavelengths so it was assumed that input features have a level of spatial continuity and hence used 1-D CNN for their analysis. Dataset was randomly divided into test set and training set in the ratio of 10 % and 90% respectively and 5-fold cross validation was performed on training set for best results. Authors have performed minimum data pre-processing on the input features by only standardizing to zero mean and unit standard deviation across all the spectral measurements and target variables. As a result, based on the root mean squared error (RMSE) metric they observed that 1-D convolutional neural network was found to be effective but models that considered interdependencies between target variables generally performed equally well. Also, it was observed that prediction of phosphorous (P) level proved to be most difficult. *Authors have recommended future works to include auxiliary feature inputs and predicting full set of target outputs on the LUCAS dataset.*

Continuing to take the benefits of Diffuse reflectance infrared spectroscopy in visible-near (vis-NIR) range for performing soil analysis, in 2019, Padarian et al. did extensive research to compare the performance of conventional methods such as PLS regression and Cubist regression tree with deep learning-based multi-tasking CNN method. Their research was primarily focussed on to explore the use of deep learning- based CNN to predict soil properties from raw soil spectral data. Their objective was to predict soil properties from unprocessed data and avoid reduction of the high dimensionality of input features and pre-processing methods. Authors have demonstrated the capacity of the CNN to simultaneously predict 6 soil properties namely Organic carbon (OC), Cation exchange capacity (CEC), Clay particle size fraction (clay), Sand particle size fraction (Sand), pH measured in water (pH) and total nitrogen content (N) using one model. Instead of using spectrogram data in 1-D as used by Matthew et al. , authors have used Hann window , a segment length of 100 with overlap of 50 observations and a sampling frequency of one to generate the spectrograms. After generating the spectrograms, they have transformed it to logarithmic scale and generated a different 2D representation of the spectrogram data for feeding to CNN model. No Auxiliary information like land cover or land usage was used in input features. *Even though research objective was to use unprocessed spectrogram data without reducing high dimensionality but transforming spectrogram data into 2D format(function of frequency and wavelength) requires pre-processing of raw spectrogram data and it has reduced the input features from length 4200 to a matrix of 51x83 (frequency wavelength).So in my opinion it is fair to assume that data was pre-processed before feeding to CNN models and same was mentioned by authors in the conclusion section of the paper.* Authors demonstrated through their work that CNN trained in multi-task settings performed better to simultaneously predict 6 soil properties from one single model however it requires more computing time and compute processing capacity. RMSE and R^2 were used as metrics for performance measurement of all the models. In their experiments, CNN models

performed better than traditional PLS regression and Cubist models. With these results Padarian et al. were able to demonstrate that the potential of deep learning techniques is not only limited to computer vision, medical image analysis, etc. but it has potential to change the way soil spectral data is modelled.

Another study conducted by Singh & Kasana,2019 used LSTM(Long Short-Term memory) deep learning method for quantification of soil properties because hyperspectral data is sequential in nature and LSTM is preferred in solving sequential problems. Authors also addressed the high dimensionality of spectral data by applying PCA(principal component analysis) prior to feeding input features to the deep learning model. PCA is common technique used to reduce high dimensionality in time series (sequential) problems (Liu & Wang, 2018).The proposed framework by authors for soil analysis is compared against other popular models like LSTM, Partial least square regression (PLSR), Principal Component Regression (PCR), Support Vector Regression (SVR), Multiple Linear Regression (MLR) and Stepwise Regression (SWR). Authors have compared the performance of LSTM with PCA (proposed framework) and without PCA (Only LSTM) to show the effectiveness of PCA components. RMSE (Root mean square error) and R2 (Coefficient of Determination), Mean Absolute Error (MAE), Spearman correlation coefficient (r), Pearson Correlation coefficient (r) and Lin's concordance coefficient (rc).It was observed that proposed framework PCA combined with LSTM outperformed other models. The 6 soil properties which were predicted in the experiments were OC, CEC, Clay, Sand, Nitrogen, and pH (H2O).

In addition to previous studies, instead of using all the fields of wavelength , Wang, Y. et al. has performed research to identify which wavelengths (out of 4200 wavelengths) from LUCAS spectroscopic data contain more information about soil total nitrogen(STN) and then predict STN using Convolutional neural network(Inception architecture) method. Though their approach was limited to screening of wavelengths for prediction of only one soil property i.e., STN but it has provided useful insight into understanding of spectroscopic data and relationship of different wavelengths with specific soil property. Based on the experiments the authors have obtained different results for prediction of STN using both screened and full wavelength data and concluded that deep learning performed better to screen the wavelengths and as well predict the STN compared to other statistical approaches.
Table 1 presents the summary of the different research papers who did research in soil analysis using deep learning techniques.

Table 1 Summary of different research papers

No #	Paper Title	Year Published	Authors	Properties Predicted	Technique/s Used	Metrics Used For measurement	Result
1	Deep Learning Architectures for Soil Property Prediction	2015	Matthew Veres, Griffin Lacey, Graham W. Taylor	Sand, pH, Soil Organic Carbon (SOC), Calcium (Ca) & Phosphorous (P).	1-D convnet, No Auxiliary info used, 5 properties prediction	RMSE	1-d CNN performed better than other models
2	Using deep learning to predict soil properties from regional spectral data	Nov 2018	J. Padarian∗, B. Minasny, A.B. McBratney	Organic carbon (OC), Cation exchange capacity (CEC), Clay particle size fraction (clay), Sand particle size fraction (Sand), pH measured in water (pH) and total nitrogen content (N)	PLS regression, Cubist regression tree, multi-tasking CNN, spectrogram data as 2D.No Auxiliary info used. 6 properties prediction.	RMSE, R^2	Multi-CNN outperformed other models
3	Estimation of soil properties from the EU spectral library using long short-term memory networks	June 2019	Simranjit Singh∗, Singara Singh Kasana	OC, CEC, Clay, Sand, pH(H_2O) and Nitrogen	PCA with LSTM and compared results with LSTM, PLSR, PCR,SVR, MLR, SWR. No Auxiliary info used. 6 properties prediction	RMSE, R^2, MAE, ρ,r, ρc	PCA with LSTM outperformed other models
4	Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel1-D convolutional neural network	March 2020	NikolaosL.Tsakiridisa,∗, KonstantinosD.Keramarisa, JohnB.Theocharisa, GeorgeC.Zalidisb	Clay, Slit, Sand, pH, CEC, OC, CaCO3,N, P,K	1D CNN	RMSE, R^2	1D CNN outperformed other CNN models
5	Simultaneous Prediction of Soil Properties Using Multi_CNN Model	Sep 2020	Ruixue Li1, Bo Yin1,2,*, Yanping Cong1and Zehua Du1	Organic Carbon, Nitrogen, Clay	Dual-stream convolutional neural network model Multi_CNN. Combine 1D and 2D Convolution, traditional regression model	R^2	Multi_CNN performed better
6	A three-level Multiple-Kernel Learning approach for soil spectral analysis	Jan 2020	NikolaosL.Tsakiridisa,∗, ChristosG.Chadoulosa, JohnB.Theocharisa, EyalBen-Dorb, GeorgeC.Zalidisc	Soil organic carbon	Multiple kernel learning (MKL)– 3 level	RPIQ, RMSE, R^2,	3 level MKL outperformed
7	A deep learning-based method for screening soil total nitrogen characteristic wavelengths	June 2021	Yueting Wanga,b, Minzan Lia, Ronghua Jia, Minjuan Wangb, Lihua Zhenga,b,*	Soil Nitrogen	CNN-Inception model, Random Forest, Extreme Learning Method. Limited number of wavelength	RMSE, R^2	CNN performed better

3 Methodology

Software development life cycle (SDLC) is a process which is mainly focussed on design, development, and testing of software with an aim to produce a high-quality software within defined time and budget. There are various SDLC models e.g., Waterfall model, Iterative mode, V-Model, Agile model etc. and selection of an appropriate model is an important first step in any software project. After a thoughtful study of few popular models like Waterfall, Iterative, V-Model and Agile model, I have concluded on using Waterfall methodology due to following reasons:

1. **Clear Structure of the project:** Waterfall model is a linear sequential life cycle model and whole process of software development is divided into separate phases. This model is often preferred when all the phases are clear since beginning of the project. In the case of this research work, all the required stages were clear since beginning and hence waterfall model was best fit. Below figure 1 shows classical waterfall model structure but the last stage related with 'Maintenance' is excluded from the scope of the project.

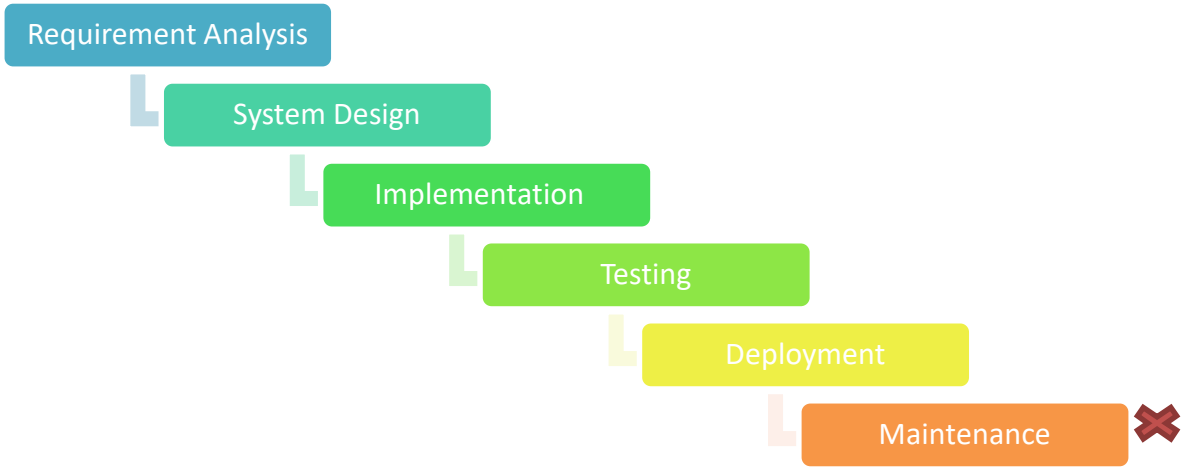


Figure 1 : Waterfall model (adopted for this research work)

2. **Clear End Goal since beginning :** One of the reasons for selection of waterfall model is early visibility of end goal and for small projects like this research work, end goals are very clear since beginning so there is very less chance of deviation from that and hence this is another reason for choosing the waterfall model methodology.

3. **Systematic approach of information flow to next stages :** Waterfall model is very methodical approach where output from the previous stage acts as an input to subsequent next stage. So, this model allows a clean flow of information across the stages. In the case of this research project, requirement act as input to system design and after completion of system design , software implementation is stared. Software development follows the system design and then its tested against the requirement in the testing phase.

4 Requirements

This section presents the technique used for finalising on the requirements for this project. Since I have selected waterfall methodology for software development as described in section 3, so it is very important that requirements were thoroughly analysed and cleared defined before moving to system design phase.

4.1 Dataset Selection

To meet the objectives of this research as described in section 1.2, it is important to select the right dataset which contains vis-NIR spectral values for different soil samples, different soil properties and auxiliary information like elevation, land coverage , land usage etc. In this regard, different websites which performed soil surveys or involved in research related with soil chemistry were explored and after an extensive search, LUCAS (Land use and coverage area frame survey) dataset was selected. This public dataset

was hosted by European soil data centre and was made available on request for study and research purposes. The dataset contained approx. ~21K sample point which is sufficient for training deep learning models.

4.2 Gap analysis

As many researchers have applied deep learning techniques for predicting soil properties in past so it was important to understand the current state of progress and available results. To understand the current progress and identify the gaps in this field, literature review was done on 8 selected research papers which were published between 2015 to 2021 in scientific journals. This gap analysis has provided insights into formulating requirements for the software development.

4.3 System requirements

Development of deep learning models and its training and validation can be computationally expensive, so it is important that a computer with a powerful configuration is selected for software development and testing purposes. For the development activities in this project, a computer with below configuration was used (Table 2).

Table 2 System Configuration

System Type	x64-based PC
OS Name	Microsoft Windows 10 Home
Processor	AMD Ryzen 7 3750H with Radeon Vega Mobile Gfx, 2300 Mhz, 4 Core(s), 8 Logical Processor(s)
Installed Physical Memory (RAM)	8.00 GB

4.4 Functionality requirements

Following are the list of functionality requirements which were considered for this project.

I. Join different data files into a single dataset before processing
II. Null value handling in dataset
III. Missing value handling dataset
IV. Handle any misalignment of data types in dataset
V. Perform exploratory data analysis
VI. Identify and remove any noise from vis-NIR spectral data
VII. Perform Data smoothening
VIII. Perform Standard scaling on dataset
IX. Perform feature dimensionality reduction using Principal Component Analysis(PCA)
X. Split the dataset into training and testing in the ration of 80%-20 % respectively
XI. Create different training and testing dataset from below combination (Table 3)
XII.

Table 3 Dataset Combinations

Dataset	Standard Scaling	PCA	Aux Information	Vis-NIR data
DS1	Yes	No	No	Yes
DS2	Yes	No	Yes	Yes
DS3	Yes	Yes	No	Yes
DS4	Yes	Yes	Yes	Yes

XIII. Create Multilayer perceptron (MLP) model with 2 different architectures
XIV. Create Convolutional Neural Network (CNN) model with 3 different architectures
XV. Train and test MLP and CNN with different training and testing dataset
XVI. Store the training and testing results into excel file

4.5 Non -Functionality requirements

I. Modular design for software
II. The software should be able to run efficiently on machine with limited computational resources as per the given system requirements.

5 Dataset

This research project will be using dataset imported from the EUROPEAN SOIL DATA CENTRE (ESDAC) website. The dataset contains topsoil information, soil spectral data, location coordinates for each sample location for 28 EU member countries including UK. The dataset is spread across many files and each observation is linked with unique pointID. Using a pointID, sample entries can be mapped across different files. There are 21,860 entries recorded in the 6 main dataset file which contain both revisited location and as well new locations added during land survey in 2015. This dataset is collected as part of the Land Use and Coverage area frame survey(LUCAS).
For the soil testing purposes, a composite sample of approximately 500 g was taken from five subsamples collected with a spade at each sample point. The first subsample was collected at the geo-referenced point location; the other four subsamples were collected at 2 m distance following the cardinal directions (North, East, South and West).

A pre-defined set of process steps were followed for the sample collection e.g., before collecting the subsamples, any vegetation residues or litter were cleared from the surface of the soil by raking with the tools. Afterwards samples were dig and kept in a bucket. The procedure was repeated at the other four subsample sites. At the end, all the five samples were thoroughly mixed with a trowel and any debris present was removed from the sample. Approximately 500 g of the mixed soil was taken from the bucket and placed in a separate plastic bag for labelling. All the soil samples were air dried first before sealing the bags and sending to central lab for detailed physical and chemical testing using standardized ISO methods (except for extractable Potassium).Table 4 and 5 shows the different attributes present in topsoil and spectral documents.

Table 4 Topsoil Attributes in dataset

Field	Description	Units/ Values	Limit of detection/ Measurement range
Point_ID	Unique identifier of the LUCAS survey point	8 digits number	–
Revisited_point	Is the point visited in both 2009/2012 and 2015 surveys	Yes / No	–
Coarse	Coarse elements	%	> 2000 µm
Clay	Clay content	%	0.02 – 2000 µm
Silt	Silt content	%	0.02 – 2000 µm
Sand	Sand content	%	0.02 – 2000 µm
pH($CaCl_2$)	pH measured in a $CaCl_2$ solution	–	2 – 10

pH(H_2O)	pH measured in a suspension of soil in water	–	2 – 10
EC	Electrical conductivity	mS/m	0.1
OC	Organic carbon content	g/kg	2
CaCO3	Carbonates content	g/kg	1
P	Phosphorus content	mg/kg	10
N	Total nitrogen content	g/kg	0.2
K	Extractable potassium content	mg/kg	10
Elevation	The elevation at which the soil sample was taken (in meter) (GPS_ALTITUDE)		
LC1	LandCover class		
LU1	LandUse class		
SOIL_STONES	Percentage of Stones in the sample, as a class		
NUTS_0	NUTS code for the country where the sample was taken		
NUTS_1, NUTS_2, NUTS_3	NUTS 1, 2 and 3 codes for the location where the sample was taken		
LC1_Desc	LandCover class description		
LU1_Desc	LandUse class description		

In addition to topsoil information , dataset also contains the spectral data of the LUCAS 2015 Soil survey (file *LUCAS2015_Soil_Spectra_EU28*). This package contained Diffuse reflectance spectra which were collected for **21782 samples** in the visible (400 to 780 nm) and the near infrared (780 to 2500 nm) regions with 0.5 nm spectral resolution. All the measurements were carried out following the protocol of the Soil Spectroscopy Group with a FOSS Rapid Content Analyzer. For every sample , two spectra were collected by rotating the sample container each time so that each measurement spots a different area of the soil sample. The values in the file are Absorbance.

The dataset package contains one CSV file for each member state. Altogether, there are 28 CSV files, including UK. In the files, each sample is identified by its Point ID and Soil ID. Each row contains a spectrum. The data are absorbance in the vis-NIR regions.

Table 5 Attributed in Spectral datafiles

Field	**Description**	**Units/ Values**
Source	Unique identifier of the sample in the laboratory (for ESDAC internal purposes only	LU_$$_$$.txt
SampleID	Unique identifier of the two spectra collected per sample	Soil ID_1 Soil ID_2
PointID	Unique identifier of the LUCAS survey point.	8 digits number
NUTS_0	ISO country code for each member state	2 letters
SampleN	Unique identifier of the soil sample	5 digits number

Spectrum between 400 – 2500 nm	Absorbance data in the vis-NIR regions (400 to 2500 nm) with 0.5 nm resolution	Absorbance

6 Design

The design of this software is divided into multiple building blocks with each block performing a set of pre-defined activities. Figure 2 shows the building blocks and their functionality. All the building blocks are interconnected to form high level design as seen in Figure 3, of high-level design. Below sections will briefly explain the functionality of each building block.

1. Block 1: Data Loading & EDA : This is starting design block in the workflow and is mainly responsible for reading both topsoil and spectral data input data files into memory and performing exploratory data analysis (EDA). The output from this block provides insight into data attributes, data sanity checks, data distribution, outliers in the data and visual summary of EDA. The insights gathered from this block acts an input to next block which deals with data pre-processing.

2. Block 2: Data Pre-processing : Data pre-processing is one of the key steps in any machine learning project. Often the input data is not in the desired format for directly feeding to machine learning models and hence needs to undergo different transformations before it can be used for model training. In addition to data cleaning and transformations like data smoothening, standard scaling, feature dimensionality reduction, conversion from categorical to numeric for machine learning, splitting of input data into training and testing subset is performed. This splitting helps to train model using training data subset and validate the accuracy of models on unseen testing subset dataset.

3. Block 3: Model Creation : This block will create a base MLP model architecture and 3 different architectures of CNN. Since multiple combination of training and testing data subsets would be used with each architecture model, so pipeline would be created for easy experimentation purposes. Also, the initial hyperparameters are configured for each architecture in this block of execution.

4. Block 4: Model Training : This is the execution block of the design and different combinations of training and testing data subsets are fed to neural network pipeline. Models losses are monitored in this step and passed as input to next block for model evaluation.

5. Block 5: Model Evaluation & Results: The design block is model evaluation which mainly focuses on comparing losses from each model training and testing and use that information to fine tune the model hyperparameters e.g., Number of hidden layers, dropout layers, batch normalization, kernel size, max pooling, and activation functions. This block also ensures that results are captured and saved on disk for future analysis.

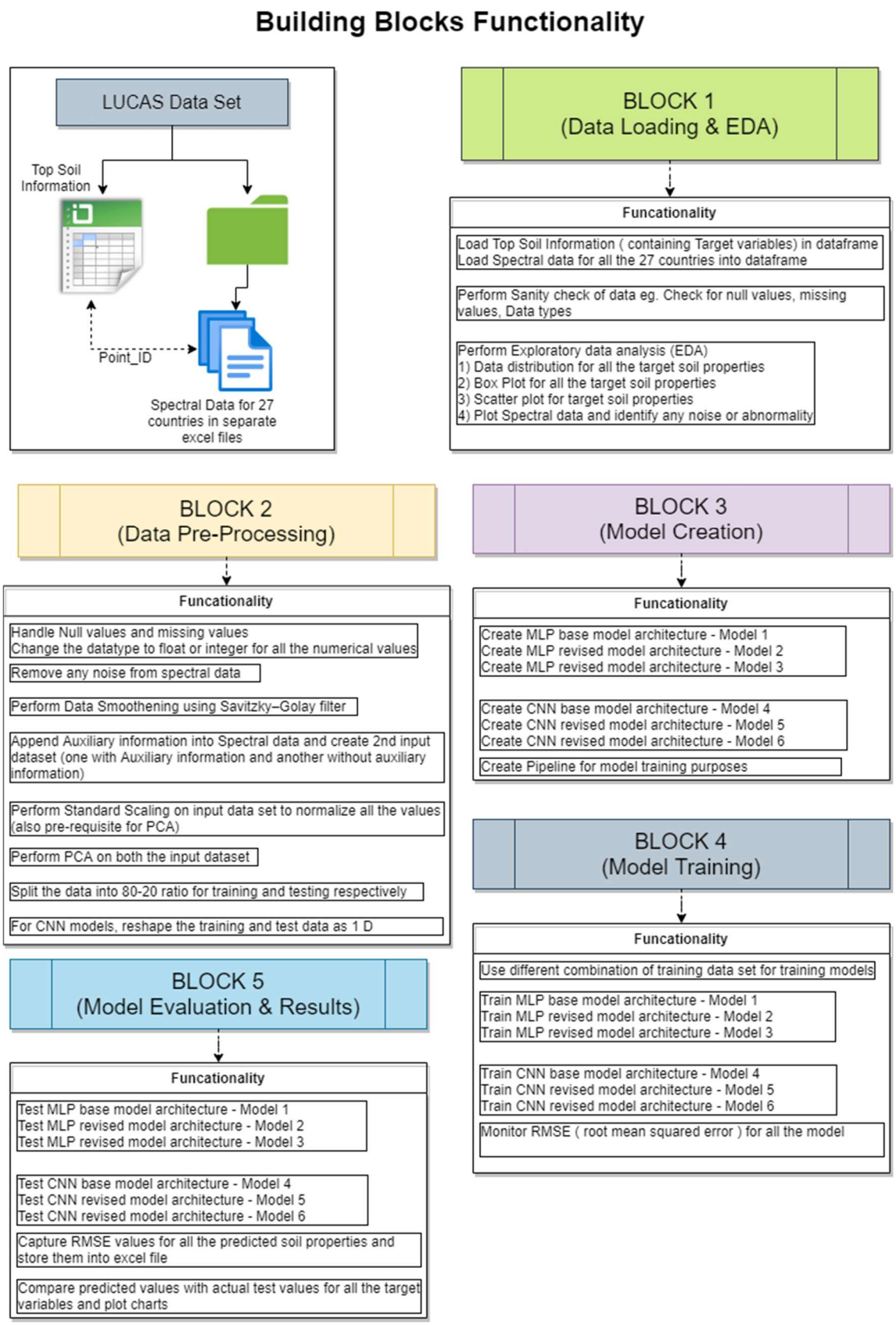


Figure 2: Design building blocks

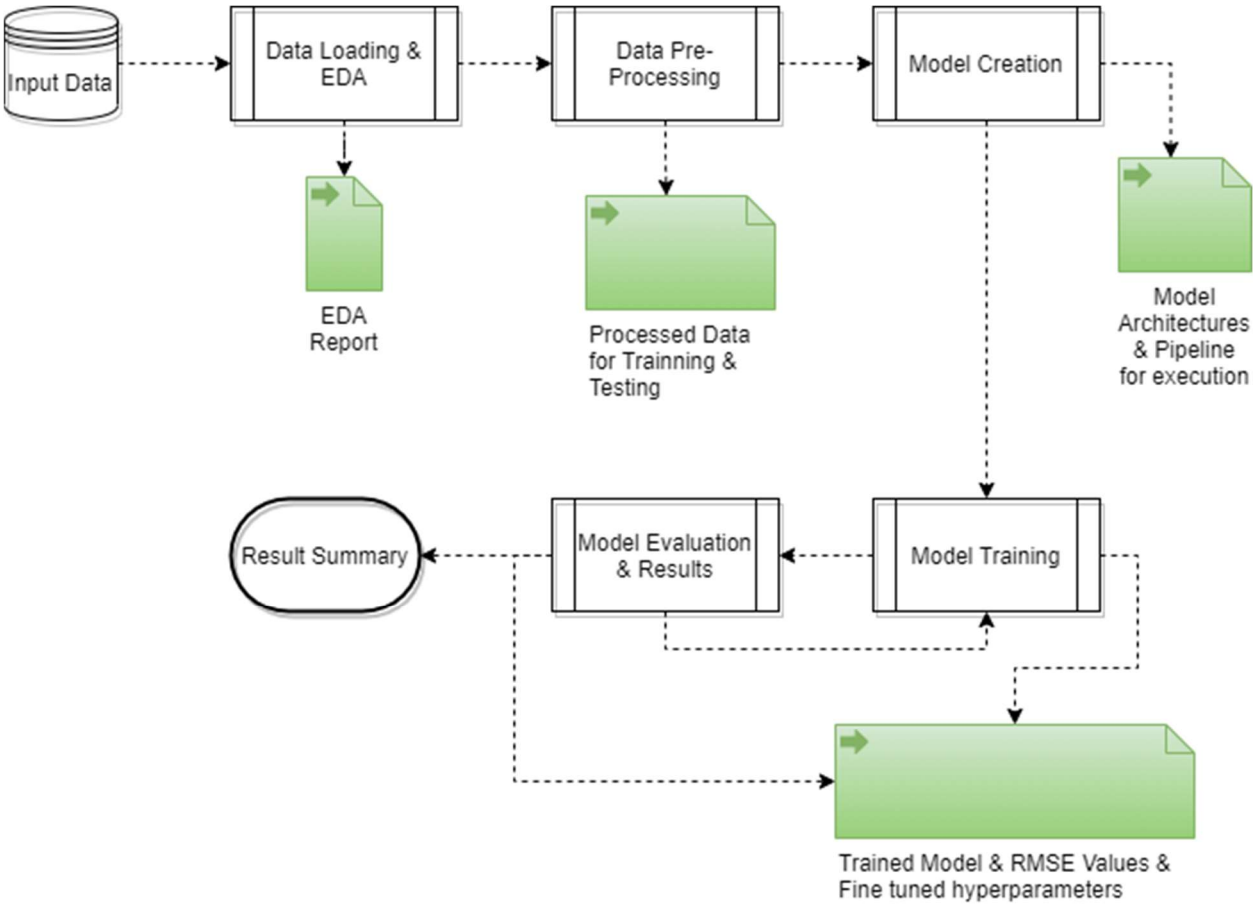


Figure 3: High level design

7 Techniques

7.1 Feature Scaling

It is important to rescale all the features value when the dataset contains values of different scale otherwise machine learning based algorithms training will get biased with high values in the feature set. Mainly 2 methods are used for scaling (Bhandari, A. ,2020).

1. *Normalization* method– This method limits the rescales and fit all the values between 0 and 1. It is also referred as Min-Max scaling method because it has minimum and maximum limit for each value.

2. *Standardization* method–Compared to normalization where values are limited between 0 and 1, standardization replaces each value with its distance from the mean in unit of standard deviation. Equation (1) refers to formula for calculating standardization.

$$x' = \frac{x - \mu}{\sigma} \qquad (1)$$

Also, feature standardization is pre-condition in case dimensionality reduction is applied on the dataset.
In this research, standardization is preferred method for feature scaling because it does not limit values between 0 and 1 and is performed prior to dimensionality reduction.

7.2 Data Smoothening : Savitzky-Golay

This section discusses about data smoothening technique, most widely used Savitzky-Golay low pass filter for removing unwanted fluctuations from the data sample points. This is often performed on noisy signals to obtain a signal that is easier to analyse and understand. This low pass filter achieves this smoothening of noisy data by fitting a polynomial to a set of input points and then evaluating the resultant polynomial (Savitzky and Golay ,1964).

7.3 Feature Dimensionality Reduction : Principle Component Analysis (PCA)

Dimensionality is described as the number of available feature vectors in the dataset. Due to high dimensionality of the dataset, the computational cost of the training model increases significantly and sometime the information present in dataset is highly correlated which does not provide any new insight hence it is required to reduce the feature dimensionality. To address the problem of high dimensionality, some popular techniques like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) etc. are applied on dataset and data is transformed into smaller feature set without significantly losing any information. In this research, PCA is applied for reducing high dimensionality.

This method works by creating Principal components which are new features, created by combination of one of more variable from the dataset. This technique calculates principal component by using eigen vectors and eigen values of the covariance matrix. The goal of the PCA is to combine maximum variant information i.e., largest possible variance into first component and then maximum remaining into second and so on until there is no change in variance (Jaadi, Zakaria ,2020).

7.4 Multi-Layer Perceptron (MLP)

Multi-layer perceptron (MLP) is a supplement of feed forward neural network. There are mainly 3 types of layers in any MLP. First layer is usually called as the input layer, last layer is referred as output layer and layers in between the input and output are hidden layers, as shown in Figure 4.

The input signal is fed to input layer for the processing. The output layer can either perform classification task or regression as per the requirement. The hidden layers between the input and output layers perform maximum computation in the case of MLP and are considered as and engine of the neural network. In the case of MLP, dataflows in the forward direction from input to output layer like feed forward network. MLP uses back propagation learning algorithm in training the neurons. MLPs are found suitable to approximate any continuous function and are used in solving problems where outputs are not linearly separable (Abirami and Chitra, 2020).

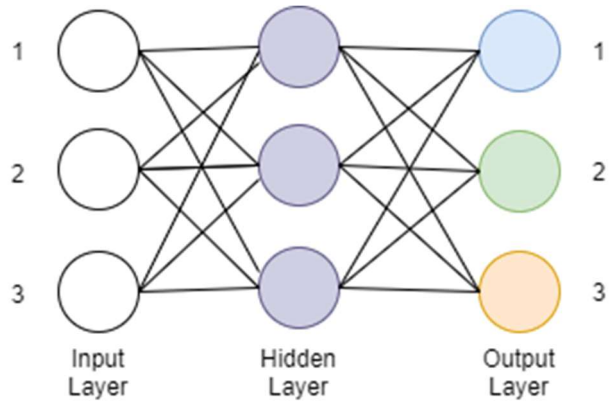


Figure 4 Multi-Layer Perceptron (Abirami et al.,2020)

Below equations shows the calculation performed in each neuron of hidden and output layers :

$$o(x) = G(b(2) + W(2)h(x)) \quad (2)$$
$$h(x) = \Phi(x) = s(b(1) + W(1)x) \quad (3)$$

where b(1), b(2) are bias vectors ; W(1), W(2) are weight matrices and G and s are activation functions.
Rectified Linear Unit (ReLU) is the preferred activation functions in the intermediate hidden layer and Sigmoid function in the case of classification problem and Linear regression in case of regression problem are mostly used at the output layer.

7.5 Convolutional Neural Network (CNN)

Convolutional neural networks or CNNs are a specialized type of neural network for processing data that has known grid-like topology e.g., time-series data(1-D grid of samples at regular time interval) or image data(2-D grid of pixels). Since network uses mathematical operation called **convolution** instead of general matrix multiplication in at least one of its layers hence its named as convolutional neural networks.

Convolution operation is usually denoted with an asterisk. First argument to the convolution (in this example x) is referred to as **input** and second argument (in this example, the function w) as the **kernel.** The output is sometimes referred as the **feature map.**

$$s(t) = \int x(a)w(t-a)da \quad (4)$$
$$s(t) = (x * w)(t) \quad (5)$$

Convolutions are often used over more than one axis at a time but in the case of 1D CNN ,there is only 1 dimension and data is used as a continuous time series data.

There are 3 important characteristics of Convolution neural networks (CNNs) which provide them edge over other machine learning algorithms:

1. Sparse Interaction
2. Parameter Sharing
3. Equivalent Representations

Figure 5 depicts the typical CNN layer architecture.

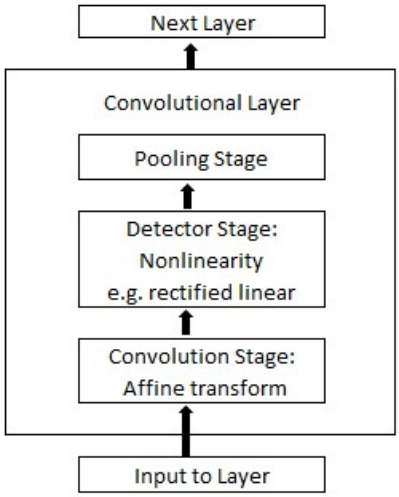


Figure 5 Typical CNN layers (Goodfellow et al., 2021)

In the first stage , layer performs several convolutions on the input data in parallel to produce a linear activations. While in second stage each linear activation is passed through a nonlinear activation function such as ReLu (rectified linear unit).It is also called as **Detector stage.** And in the final third stage, a **pooling function** is used to modify the output of the layer further (Goodfellow et al., 2016).

8 Implementation

This section will discuss about software implementation for the project. All the code is written using Python programming language and maintained in GitHub repository.
For data loading and initial pre-processing, python's Pandas package is used along with scikit-learn/SciPy machine learning software packages and all the graphs are created using matplotlib and seaborn libraries. Finally, the required libraries for creating deep neural network are imported from TensorFlow and Keras packages of deep learning.

8.1 Data Smoothening

The library for implementing Savitzky-Golay filter was imported from SciPy-signal software package.

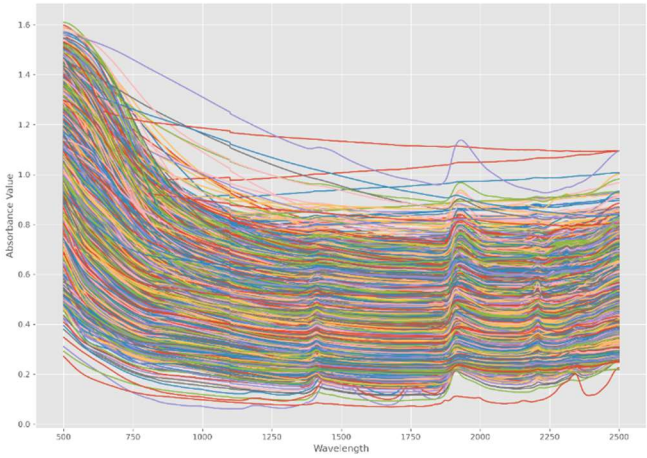


Figure 6 shows the full spectral data between 500nm – 2500nm wavelength. As it can be there is very wide range of value distribution on the y-axis and high fluctuations can also be observed. It's a indicator that there may be a presence of noise in the signal which could potentially impact the predicted results for the soil properties.

Figure 6 Before applying Savitzky-Golay filter

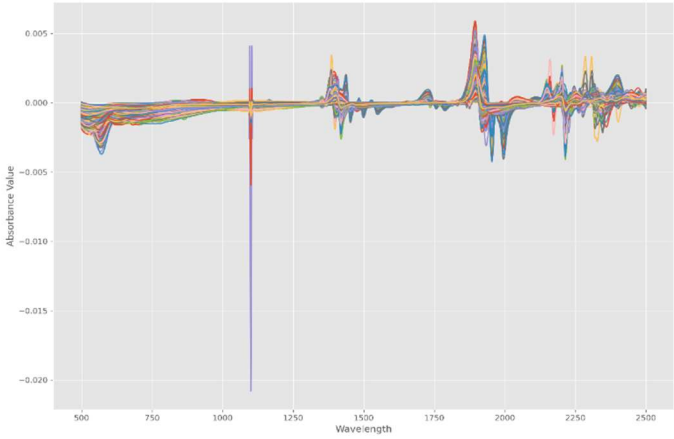


Figure 7 shows the plot of spectral data after applying the low pass Savitzky-Golay filter. As can be seen from the output plot the range of variation on y-axis is reduced to a greater extent however there is a big spike present in the middle of the data.

Figure 7 After applying Savitzky-Golay filter

8.2 *Principal Component Analysis (PCA)*

PCA was implemented in python using scikit-library package with a variance set at 95%. Explained variance and Cumulative explained variance which explains the variance of data by the principal components are plotted and shown in below figure 8 and 9 respectively

Before applying PCA, number of feature vectors in input dataset was 4200 whereas after applying PCA it was reduced to only 17 principal components. PCA has significantly reduced the dimensionality of input dataset.

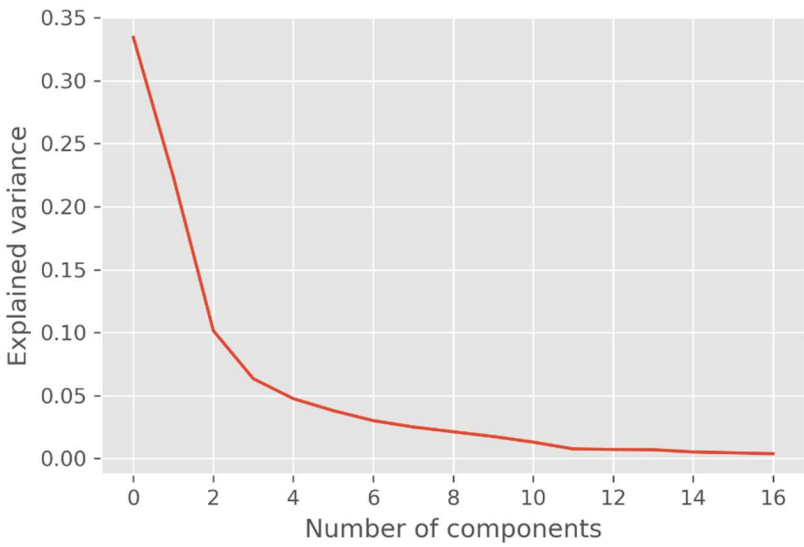


Figure 8 Explained Variance

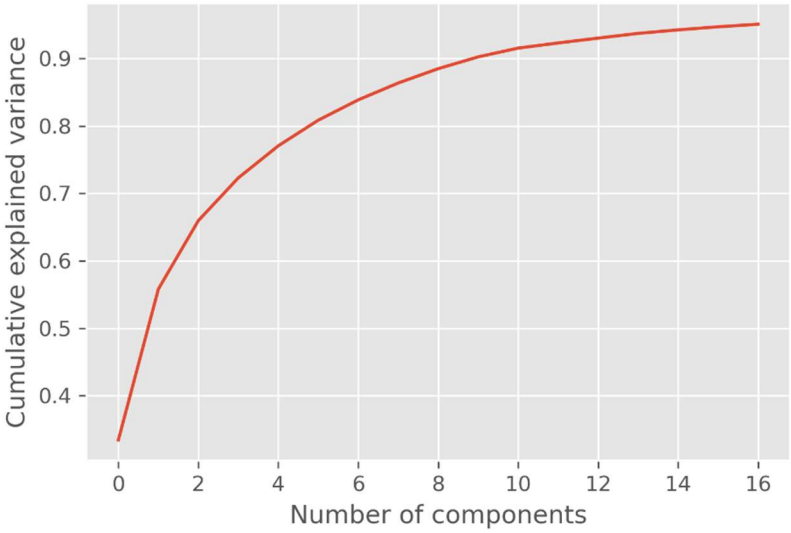


Figure 9 Cumulative explained variance

8.3 MLP Architecture

Figure 10 shows the representation of multi-layer perceptron network with a input layer, hidden layer with 50 neurons and a output layer with 8 target variables which are predicting different soil properties. The implementation of MLP neural network is done in python using TensorFlow and Keras deep learning software package libraries.

In the case of MLP, only one base architecture is designed, and different combination of input dataset is trained and tested on the network. Data splitting is performed into training and testing data subset with 80-20 ratio respectively.

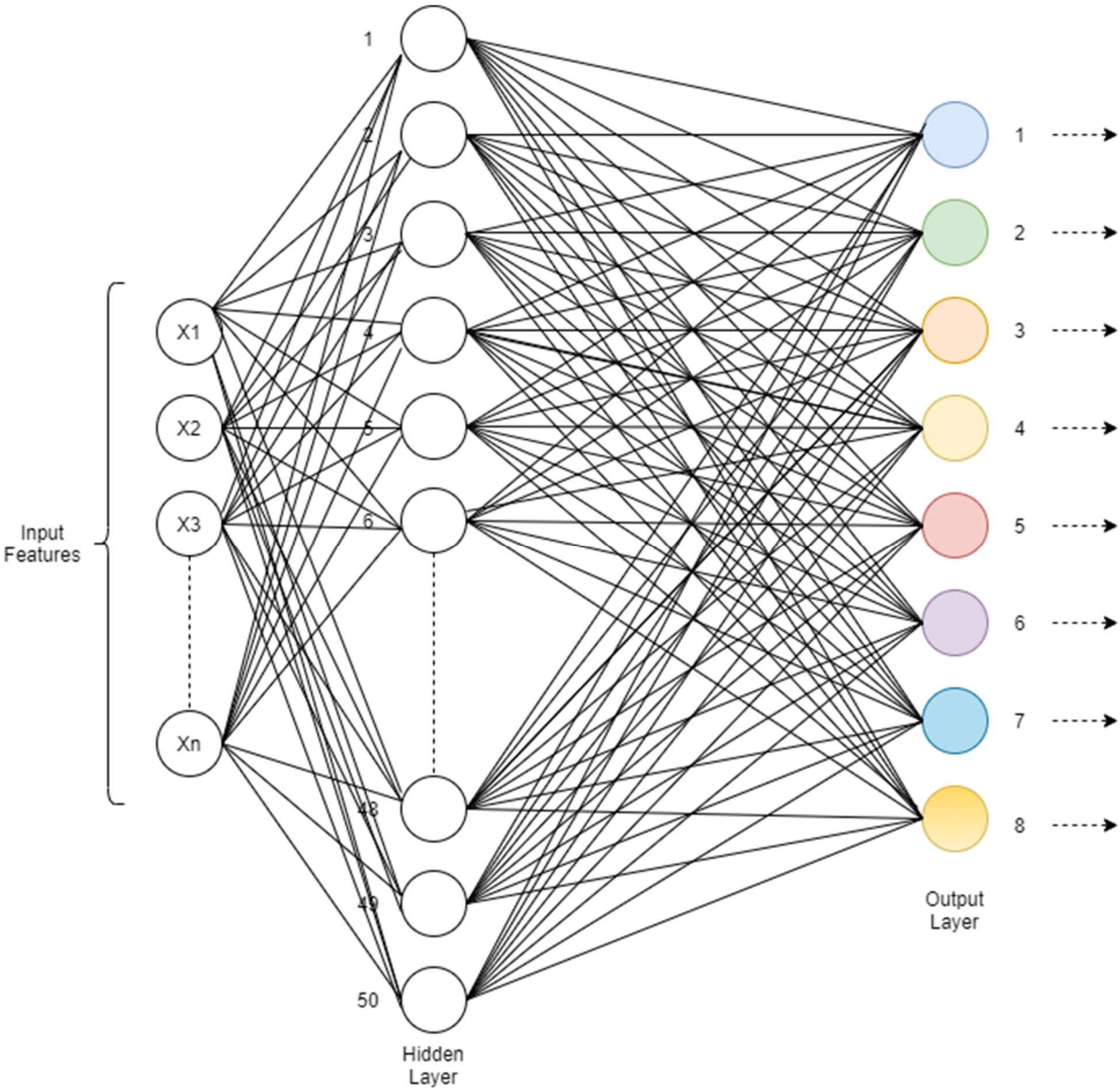


Figure 10 MLP Base Architecture

8.4 CNN Architecture

Figure 11 shows 3 CNN architectures with different number of layers , filter size, kernels and then all the 3 architectures are trained and tested using training and test data subsets. Like MLP, the input data is split into training and test in the 80-20 ratio respectively.

Mean Squared error function was selected as a metric for measure loss because it is best suited to calculate the deviation of predicted values from the actual values in the case of multi – output regression problems.

Dropout layers are used strategically in between the layers after convolution and batch normalization to avoid overlearning of the features and causing trained models to be overfit.

In the case of CNN, the dataset which was processed by applying PCA to reduce feature dimensionality was selected as input to expedite the training and testing of the models. However, data sets contained combination with and without auxiliary information to determine the impact of models prediction if additional auxiliary information added any extra benefit in prediction.

The implementation of different CNN models was done in python by using TensorFlow and Keras deep learning software packages.

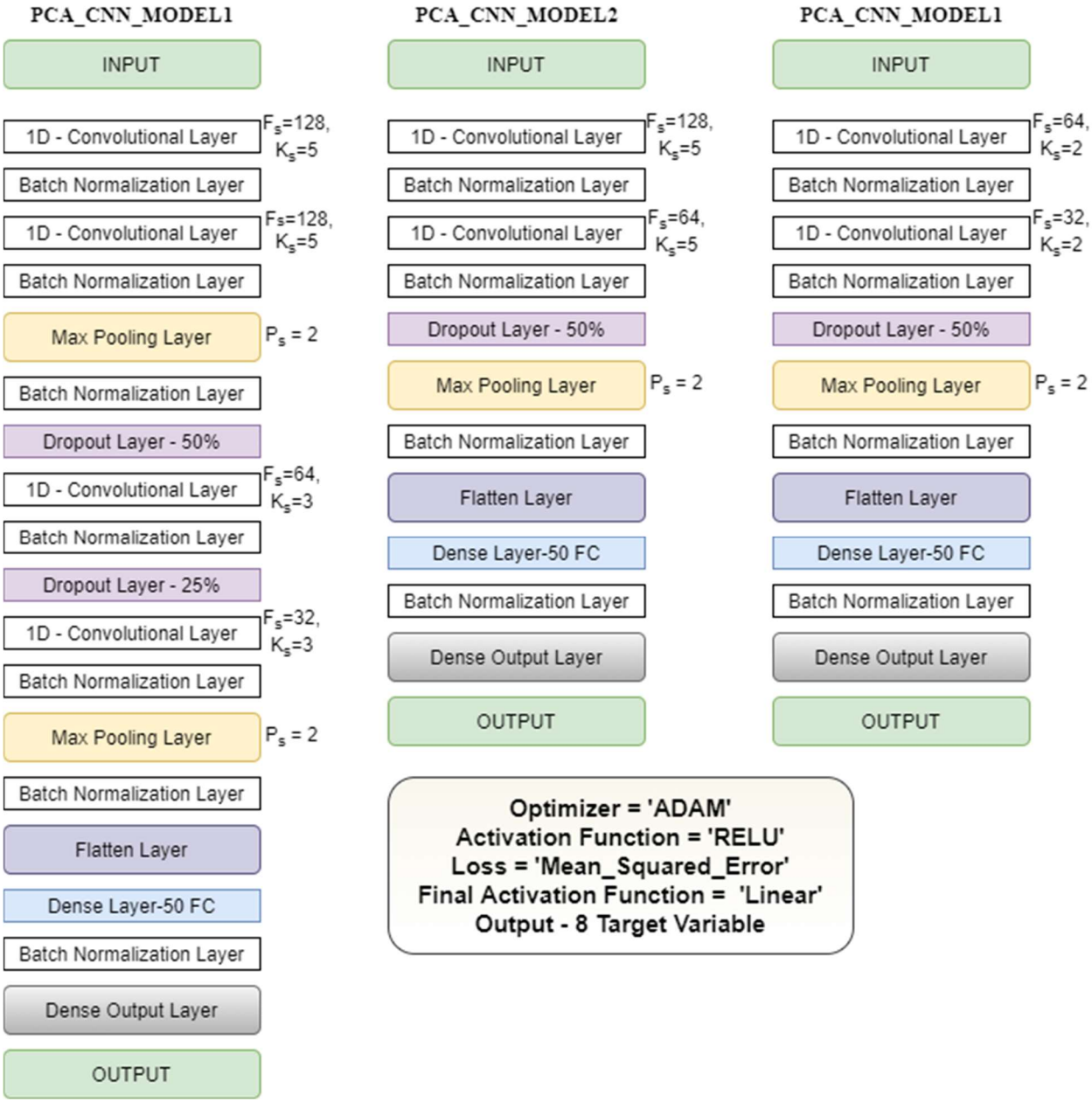


Figure 11 CNN Architectures

8.5 Metrics : Root Mean Squared Error (RMSE)

Root mean squared error (RMSE) is the square root of the mean of the square of all the errors. Error is calculated as the difference in prediction value and the actual value. The use of RMSE is very common and is often considered as error metric for measuring performance in regression analysis (Neill & Hashemi, 2018). The formula for calculating RMSE is shown below :

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2}$$

where *Oi* are the observations, *Si* predicted values of a variable, and *n* the number of observations available for analysis.

9 Results

This section is divided into 3 sub-sections and discusses the results of Exploratory data analysis, prediction error values obtained from MLP and CNN results respectively.

9.1 Exploratory Data Analysis(EDA)

Exploratory data analysis is mainly performed on raw dataset to extract information as it is from the dataset. Figure 12 shows 8 target variables and their respective values range using box plot diagram. This is done to find if the available target value has any outliers present which in turn may impact our training of models. As seen from the below box plot diagram, Potassium target value seems to contain maximum number of outliers and we can expect maximum prediction error in its prediction. pH(CaCl2) and pH(H2O) are observed as most accurate, and we would expect its prediction to be also accurate from the deep learning models.

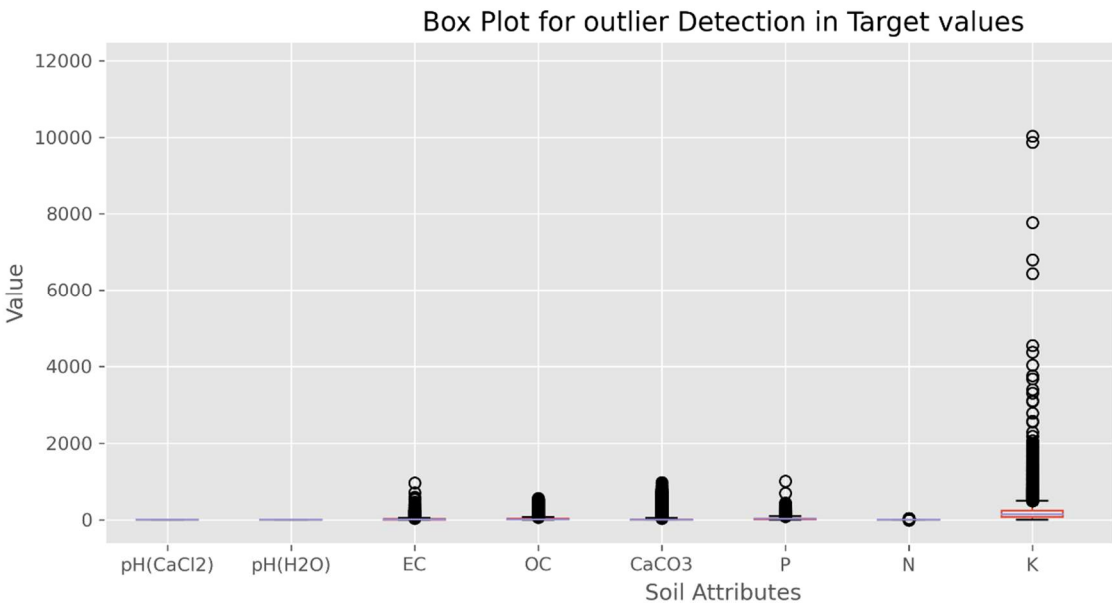


Figure 12 Outliers in target variable

In addition to finding outliers in target values, I have also represented heatmap of each of target variables in figure 13. This heatmap gives a visual representation of values distribution across different countries in their respective scale with different colours. Heatmaps are good way to show contrast in values across different sample locations for a particular soil property and it helps to understand soil chemistry across different geographical location.

Figure 14 uses histogram to show data distribution of all the 8 target variables. It is another way of looking at data to understand the distribution of different values for a particular soil property. For example, we can see from the histogram data that pH(CaCl2) & pH(H2O) sample values are uniformly distributed and maximum concentrated in the range of 4 to 8 whereas in the case of Phosphorous, maximum samples are near the left side of the histogram and could be left screwed. Similar is the case for remaining attributes like EC, OC, K, N, and K. Elevation attribute is only plotted for reference and is not actually a target variable.

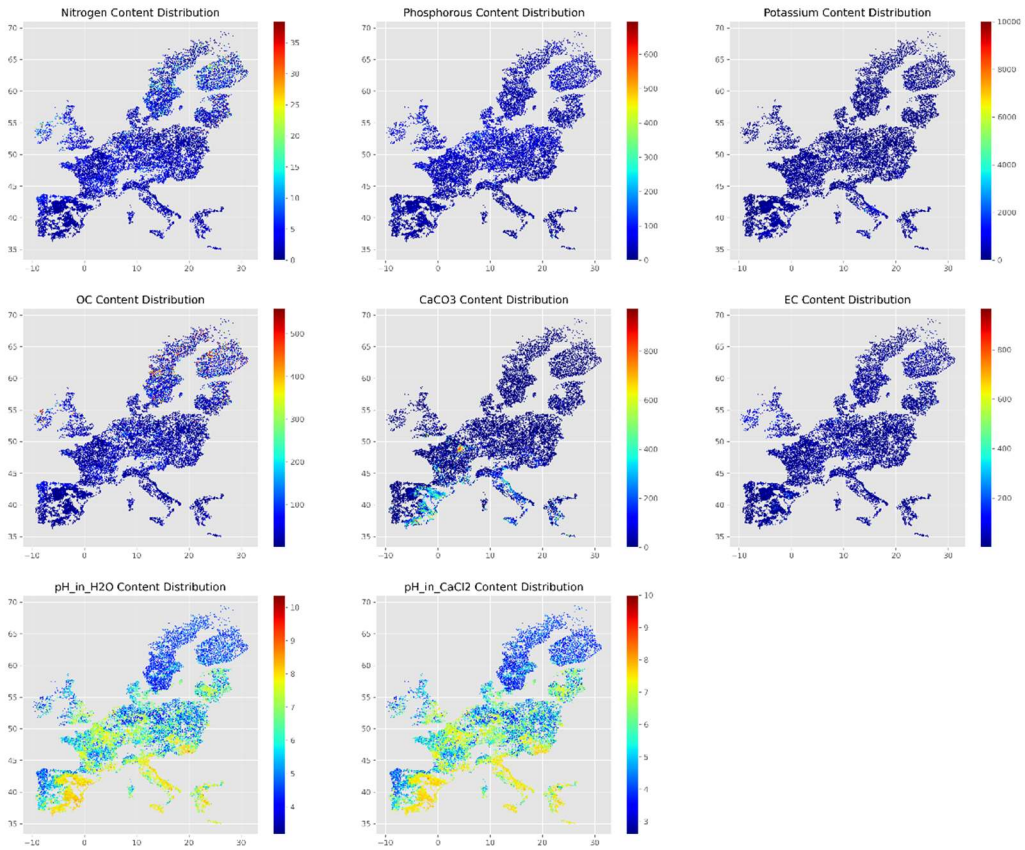


Figure 13 Heatmap of different soil attribute

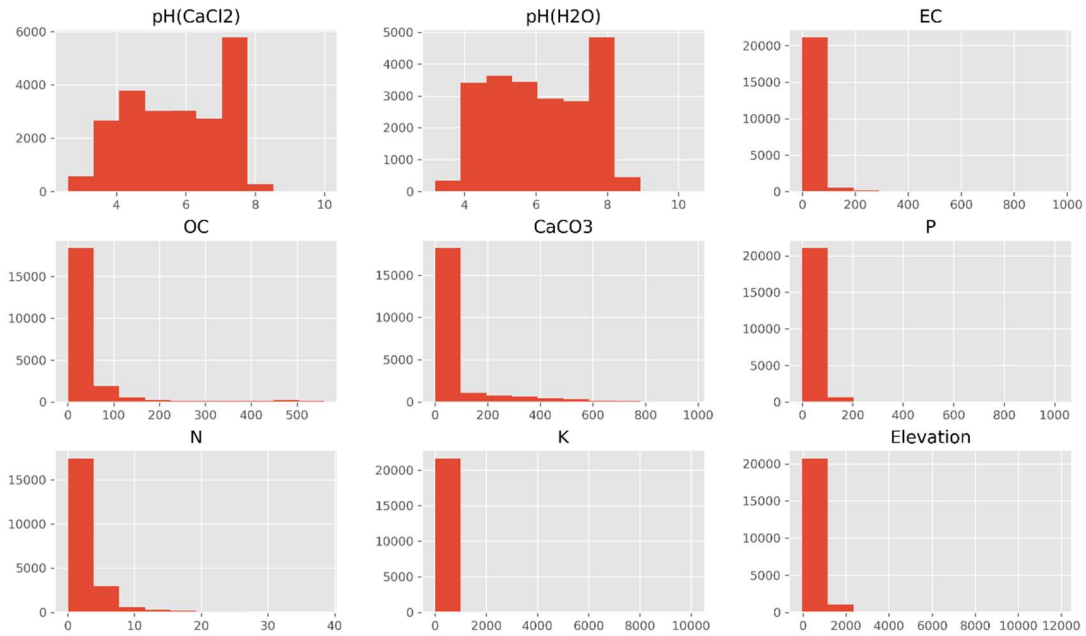


Figure 14 Data distribution of target variables

Figure 15 and Figure 16 are plots of input vis-NIR spectral data for all the sample points before and after noise removal respectively. As seen in figure 15, there is spike in curve between wavelength of 400nm – 500nm and indicate abnormal behaviour because there cannot be such high spike between neighbouring wavelengths and could be a sign of noise presence. Hence noise removal was performed by deleting those data values between wavelength of 400nm to 500nm.

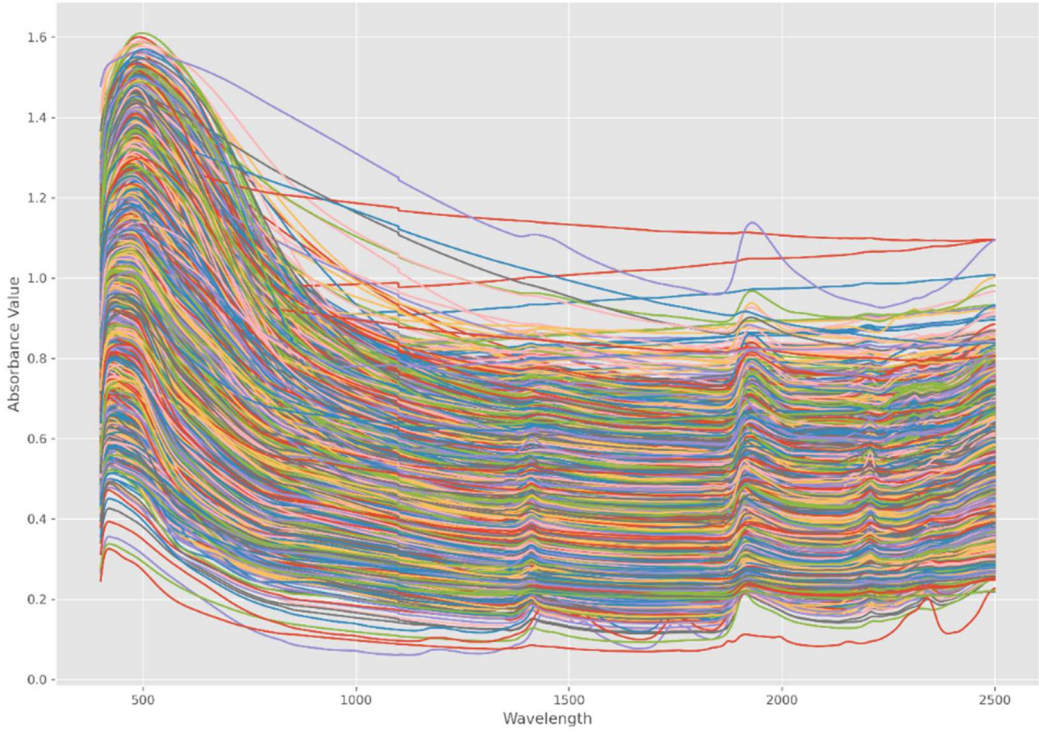


Figure 15 Original Spectral Data

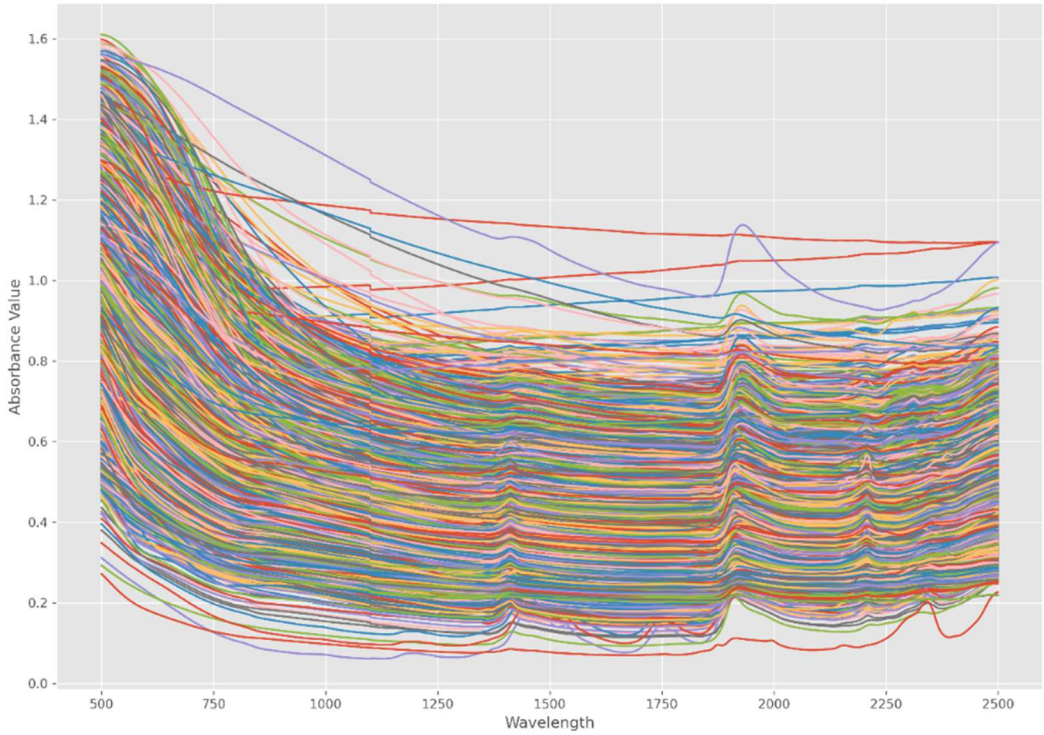
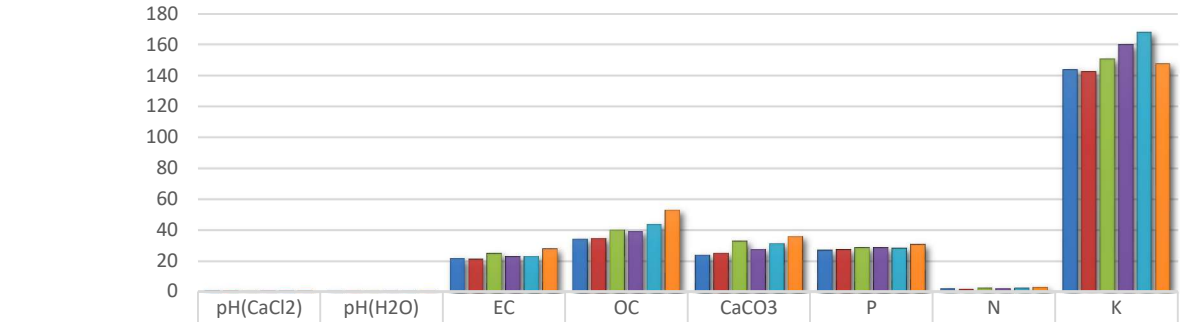


Figure 16 Spectral Data After Noise Removal

9.2 *Deep learning Models (MLP and CNN)*

6 different experiments were performed on base MLP and 3 different CNN architectures using different combination of input dataset. Root Mean Squared Error (RMSE) was used a loss function for measuring the prediction error. The goal was to keep the RMSE value to minimum or as close to 0 for accurate prediction. Figure 17 presents the result summary for all the experiment table in the plot shows the RMSE values in each case. Dataset was split into training – testing subset in the ratio of 80-20 % prior to training and testing.

1st experiment was done on MLP using the original dataset with the noise removal. No auxiliary information was added into the input data and simultaneous prediction of 8 soil properties was performed.
2nd Experiment was repeated on MLP again but with a modified dataset which contained noise removal and auxiliary information like elevation, land cover and land use.
3rd Experiment was again using the MLP model but with a further processed data which contained noise removal, auxiliary information and PCA was performed for feature dimensionality reduction.



	pH(CaCl2)	pH(H2O)	EC	OC	CaCO3	P	N	K
MLP	0.966125357	0.842434851	21.86503779	34.46632906	23.98962928	27.02924159	2.192581435	143.7444425
MLP_AUX	0.875827349	0.86027122	21.22249634	34.74007311	25.19271556	27.48534559	1.936280578	142.6193454
MLP_AUX_PCA	0.851927642	0.831723092	25.04704494	40.29576924	32.97637505	28.87580051	2.538160567	150.6958797
PCA_CNN_Model1	0.885733807	0.837520538	22.95080578	39.41037356	27.73889922	28.71634734	2.382897227	159.9078695
PCA_CNN_Model2	0.871997604	0.81202787	23.26400773	43.85338172	31.48134846	28.4662431	2.557247323	167.906383
PCA_CNN_Model3	0.883796993	0.828961749	27.93881879	53.10998173	36.14608813	30.8809326	3.187622297	147.4490769

MLP MLP_AUX MLP_AUX_PCA PCA_CNN_Model1 PCA_CNN_Model2 PCA_CNN_Model3

Figure 17 Result RMSE Summary

4th Experiment was carried on CNN model which contained 18 layers as explained in section 8.4 and shown in Figure 11. All the 3 CNN models have data with Noise removal , Auxiliary information and PCA performed to reduce feature dimensionality.

5th Experiment was repeated on second CNN model with 11 layers as explained in section 8.4 and shown in Figure 11 and same dataset as experiment 4th was used.

6th Experiment was repeated on third CNN model with 11 layers but with different filter and kernel size as explained in section 8.4 and shown in Figure 11 and same dataset as experiment 4th was used.

Key observations from all the experiments are as below :

- Error in prediction of pH(CaCl2 & H2O) & Nitrogen(N) is lowest compared to other properties
- Highest error is seen for Potassium(K) property (probably due to outliers in input target values!)
- Very less difference in prediction error for Phosphorous(P) across all the models
- EC prediction error is observed to be lowest for model using additional Auxiliary information & w/o PCA
- OC/CaCO3 prediction error is observed to be lowest & close for MLP models using with & without Auxiliary info but without PCA

10 Project Management

In general context project management covers all the aspect of the planning, delegating, monitoring and control of the project, to achieve the project objective within defined time and budget with expected quality. Project management is very important to keep the project on track and ensure successful outcome. In the following sub-sections, three main aspects of project management i.e., Schedule, Risk and Quality management are covered regarding this research project.

10.1 Project Schedule

Project timeline and clarity on the work items are very important for success of any project. In this regard, work breakdown structure was used to create list of activities and based on the estimates timelines were assigned in the beginning of the project. Excel was used to create a Gantt chart swim-line view for all the activities.
Initially it was planned to start the report creation in parallel to coding and execution but due to complexity in model training because of computational issue, I must delay working on the report part and more time was dedicated to coding and testing of models.
Detailed project schedule with work breakdown and schedule timeline is provided in appendix A for reference.

10.2 Risk Management

Risk management is the process of project management that mainly focuses on managing risks occurred during the project execution. Risk is defined as unforeseen event which can have positive or negative impact on the project outcome, so it is very important to manage them during the project execution.
Below table shows the risk register which contained potential 3 risks during the project and its mitigation plan.

Sl No.	Identified Risk (Potential risk which can occur during the project)	Impact (What is the impact due to risk)	Risk Severity (based on impact)	Mitigation Plan (What is action plan if the risk occurs)	Materialized
1	Any Health issue due to Covid-19	Could lead to delay in project execution and submission	High	Request for Extension if any health issue in the last phase of the project	No
2	Need of computation power to process huge dataset	Could take long time to train models and fine tuning which could impact on delivery timelines	High	Split data into manageable size and also try dimensionality reduction techniques to reduce data features	Yes
3	System Crash	Loss of source code and report leading to high impact on delivery timelines	High	Regular backup of data in SharePoint and local machine	No

10.3 Quality Management

Quality is another important aspect of project management activities which needs to be managed to produce an acceptable output at the end of the project.
In this project, design blocks were created for different set of functionalities and high-level design was followed since beginning of the code implementation to ensure the code is modular and scalable. Also, regular feedback was taken from supervisor to ensure the output is of good and acceptable quality as project outcome.
Project schedule was used as guidance for work activities and timelines and plan was updated on weekly basis to ensure all the tasks are covered in timely manner.

10.4 Social, Legal, Ethical and Professional Considerations

In machine learning projects, most often the social , legal, and ethical issues are raised due to the usage of dataset which contains sensitive information and can lead to individual. In this project, the dataset was sourced from public website and was freely available for study and research purposes so use of dataset has not raised any ethical or legal concerns.
Throughout this report, appropriate sources are cited and given due credit if any information from their research is used for any purposes in the research.

11 Conclusions

To summarize the research project findings, below key points are noted:

- Out of all the models, MLP models performed better for many properties compared to complex CNN models.
- PCA has helped to reduce the computational complexity by reducing feature dimensionality and model performance is at par with other models without PCA.
- Simple CNN model have shown better performance than complex CNN model.
- Complex CNN models have not shown significant improvements over simple MLP models which indicate that may be CNN models are over complex for this task.
- Presence of Outliers in the target variables during training can led to high error prediction as seen in case of Potassium(K) property.

11.1 Achievements

Regarding the first objective of this research work, “Can deep learning technique accurately predict multiple soil project simultaneously” , I believe it’s fair to say that this research has carried out methodical approach to demonstrate that deep neural networks are capable to predict multiple soil properties using a single model and it is not necessarily needed to have multiple models for predicting different soil properties.
In addition to this, the second objective ‘ Does auxiliary information helps to predict better’ also was covered as part of the research experimentations and its observed that the model prediction error rate were reduced for few properties on addition of those extra information however it was not true for all the soil attributes. So, in my opinion there is a need of further research on this objective to find the reason why only few soil properties prediction error rate was reduced on addition of auxiliary information and not every soil property.

11.2 Future Work

In every research project, there is always possibility of improvement and its encouraged to have recommendation for future work. From this research, its observed that some soil properties like pH (H2O) and pH(CaCl2) have achieved very low prediction error rate but others have not. It suggests that there is possible strong relationship between soil properties and range of wavelength. It may be possible that certain properties are more prominently identified in specific wavelength range and may be entire range of wavelength data is not required for model training. So, It is recommended to study the relationship between vis-NIR spectra wavelength ranges and different soil properties. Based on this new data subset containing the common range of wavelength which contain overlap of all the properties can be created and then prediction error can be compared for different soil properties against the original dataset containing data of all the wavelength.

Also, It would be interesting to try out more complex CNN model architecture using transfer learning. I believe this field of soil analysis is still in early stages of applying deep learning techniques for analysis and more research is needed.

From the discussions with respected supervisors, it is also suggested that since soil samples contain mix from different depths so input spectral wavelengths might not accurately represent soil properties and ‘spline’ interpolation is recommended during data pre-processing in future work. This may help in reducing the error predictions.

12 Student Reflections

This research project has been quite challenging and at the same time rewarding in terms of learning new subject areas and applying my course learnings into solving a practical problem which could be useful in soil analysis and mapping. To create a good deep learning predictor model, it was important to first understand the soil chemistry to interpret the input dataset and identity good and bad data.

Project management and in particular time management has been another secondary important learning outcome from this research project. Due to covid-19 restrictions, it was difficult to keep self-motivated and at the same time keep the progress of the project as per the plans. I had some deviation during the middle of project due to unforeseen challenge related with computational power need to train the model with huge dataset. But this problem gave me an opportunity to find an alternate way to manage model training with smaller subsets of data and helped me in demonstrating the skill to manage unfavourable conditions during the project execution which is very common in real life situations.

This research project also gave an opportunity to read and review variety of literature material and helped in building good understanding of advancement in deep learning techniques and different practical use cases which otherwise would not have been possible.

I believe I had applied good techniques on the data exploratory analysis and pre-processing and am satisfied with the outcome of those. However, the error in final predictions across different deep learning models could have been lower if more time and computational power was available to try out few more experiments using some more deep learning architectures as well.
But overall, I am satisfied with the outcome of my research as I was able to answer the research questions to a certain extent and produce comparable result for some of the soil properties with other researchers.

Bibliography and References

Abirami and Chitra, 2020, Chapter Fourteen - Energy-efficient edge based real-time healthcare support system Pethuru Raj, Preetha Evangeline, et al. (Eds.), The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases, Elsevier (2020), pp. 339-368, 10.1016/bs.adcom.2019.09.007

A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Analytical Chemistry, 36 (2) (1964), pp. 1627-1639

Bhandari, A. (2020). Feature Scaling | Standardization Vs Normalization. Retrieved 10 August 2021, from https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

FAO, ITPS, 2015. Status of the World's Soil Resources (SWSR) – Main Report. FAO, Rome, Italy. URL:http://www.fao.org/3/i5199e/i5199e.pdf.

Jaadi, Zakaria (2020), A Step-by-Step Explanation of Principal Component Analysis. (2019). Retrieved 10 Aug 2021, from https://builtin.com/data-science/step-step-explanation-principalcomponent- analysis

I. Goodfellow, Y. Bengio and A. Courville, Deep learning, 2016

Li, R., Yin, B., Cong, Y., & Du, Z. (2020). Simultaneous Prediction of Soil Properties Using Multi_CNN Model. *Sensors*, *20*(21), 6271. https://doi.org/10.3390/s20216271

McBratney A.B., Mendonça Santos M.L., Minasny B., 2003. On digital soil mapping. Geoderma 117(1–2): 3–52.

Neill, S., & Hashemi, M. (2018). Ocean Modelling for Resource Characterization. Fundamentals Of Ocean Renewable Energy, 193-235. https://doi.org/10.1016/b978-0-12-810448-4.00008-2

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernandez-Ugalde, O., 2018. LUCAS soil, the largest expandable soil dataset for Europe: a review. Eur. J. Soil Sci. 69, 140–153. https://doi.org/10.1111/ejss.12499.

Padarian, J., Minasny, B., Mcbratney, A.B., 2018. Using deep learning to predict soil properties from regional spectral data. Geoderma Regional 15, e00198. https://doi.org/10.1016/j.geodrs.2018.e00198.

Singh, S., & Kasana, S. (2019). Estimation of soil properties from the EU spectral library using long short-term memory networks. Geoderma Regional, 18, e00233. https://doi.org/10.1016/j.geodrs.2019.e00233

Tsakiridis, N., Chadoulos, C., Theocharis, J., Ben-Dor, E., & C. Zalidis, G. (2020). A three-level Multiple-Kernel Learning approach for soil spectral analysis. *Neurocomputing*, *389*, 27-41. https://doi.org/10.1016/j.neucom.2020.01.008

Tsakiridis, N., Keramaris, K., Theocharis, J., & Zalidis, G. (2020). Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. *Geoderma*, *367*, 114208. https://doi.org/10.1016/j.geoderma.2020.114208

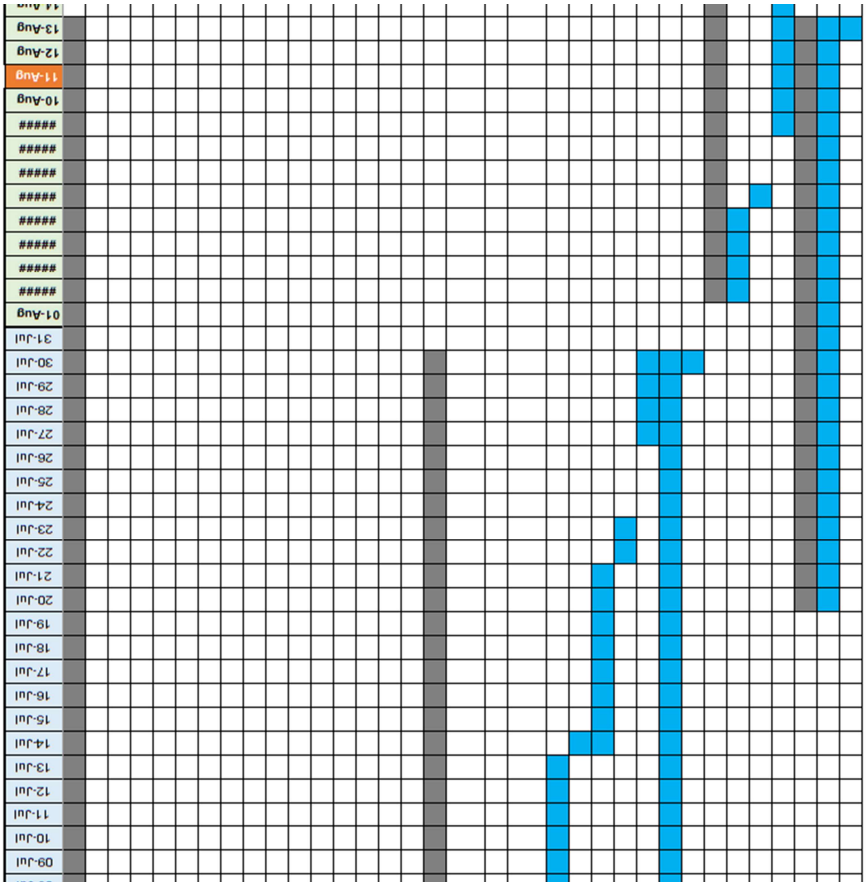
Veres, M., Lacey, G., & Taylor, G. (2015). Deep Learning Architectures for Soil Property Prediction. *2015 12Th Conference On Computer And Robot Vision*. https://doi.org/10.1109/crv.2015.15

Wang, Y., Li, M., Ji, R., Wang, M., & Zheng, L. (2021). A deep learning-based method for screening soil total nitrogen characteristic wavelengths. Computers And Electronics In Agriculture, 187, 106228. doi:10.1016/j.compag.2021.106228

Appendix A – Project Plan

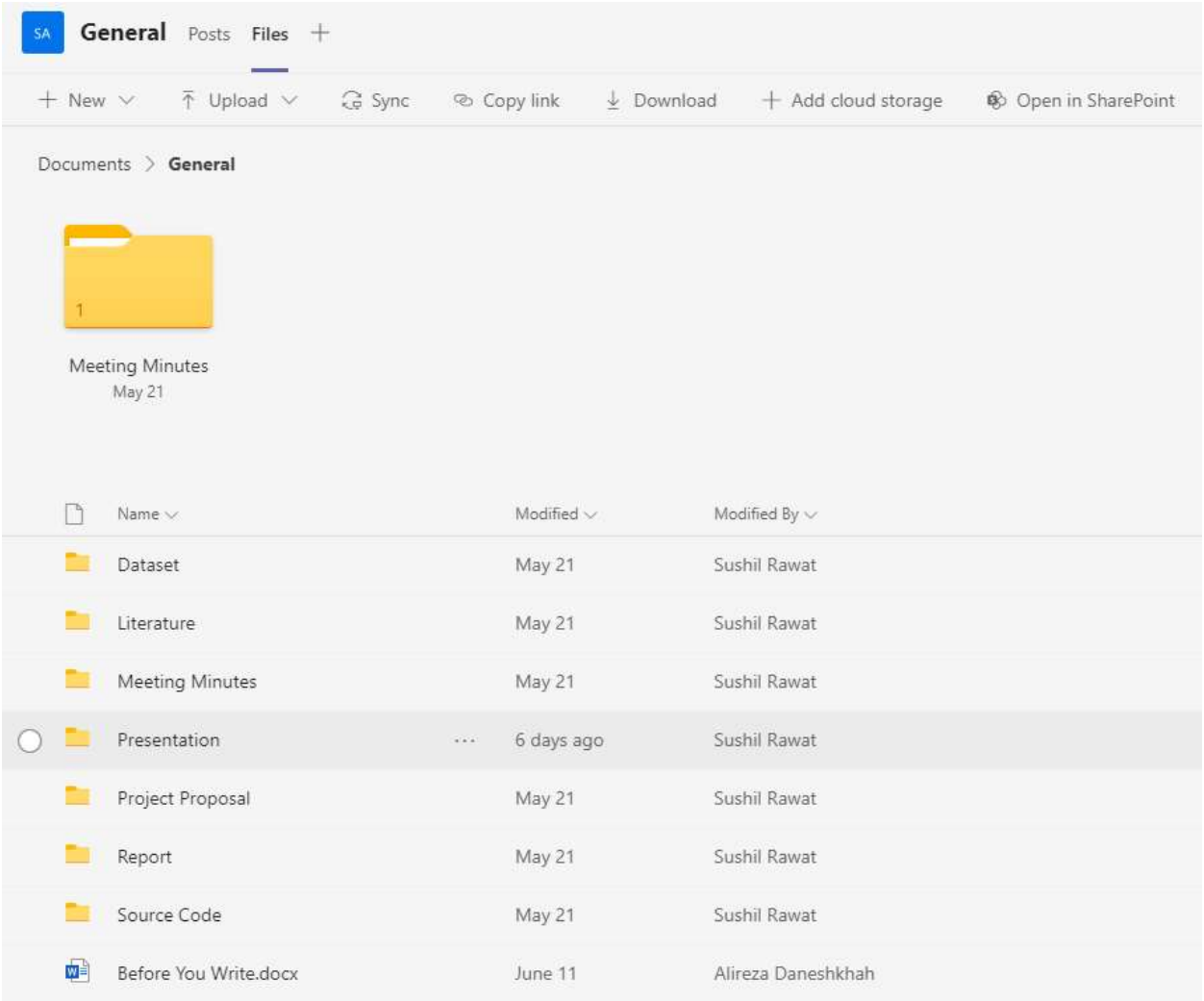
Project Name	Project Duration (Days)	Project Start Date	Project End Date	Timeline Display
MSc. Research Project	65	17-May	13-Aug	17-May

Task#	Tasks	Mileston e?	Start Date	End Date	Duration (Days)	Compl etion %
1	MSc. Research Project Plan	Project	17-May	13-Aug	65	100
2	Planning & project proposal	No	27-May	14-Jun	13	
3	Project kick-off meeting	Yes	27-May	27-May	1	100
4	Ethics Application preperation	No	02-Jun	02-Jun	1	100
5	Ethics application submission	Yes	02-Jun	02-Jun	1	100
6	Mini literature review	No	02-Jun	10-Jun	7	100
7	CW1-Submission of project proposal	Yes	14-Jun	14-Jun	1	100
8	Project Background Research & Detail Literature Review	Phase	03-Jun	21-Jun	13	
9	Identification of literature papers to review related work in the field	No	03-Jun	14-Jun	8	100
10	List best practises and short comings from the related work	No	03-Jun	21-Jun	13	100
11	Detailed comparitive study report of the literature review	No	11-Jun	21-Jun	7	100
12	Software Design Phase	Phase	22-Jun	28-Jun	5	
13	Define high level workflow for software	No	22-Jun	23-Jun	2	100
14	Define support functions needed for data pre-processing	No	24-Jun	25-Jun	2	100
15	Define communication flow between different function components	No	24-Jun	25-Jun	2	100
16	Finalize software design document	Yes	28-Jun	28-Jun	1	100
17	Software implementation & Testing phase	Phase	28-Jun	30-Jul	25	
18	Environment setup on computer for Deep learningL python & Keras library packages	No	28-Jun	28-Jun	1	100
19	Exploratory analyis of LUCAS dataset	No	29-Jun	29-Jun	1	100
20	Data pre-processing (including data wrangling, null values removal, join of environmental info into input data)	No	30-Jun	07-Jul	6	100
21	Principal Component Analysis (PCA) implementation	No	08-Jul	13-Jul	4	100
22	Data splitting for training and validation	No	14-Jul	14-Jul	1	100
23	Model training and validation	No	14-Jul	21-Jul	6	100
24	Metrics Evaluation and hyperparameters fine-tuning	No	22-Jul	23-Jul	2	100
25	UI creation and Machine learning trained model integration	No	27-Jul	30-Jul	4	100
26	Code debugging and testing	No	29-Jun	30-Jul	24	100
27	Code baseline in github repository	Yes	30-Jul	30-Jul	1	100
28	CW2 - Presenation Preperation & submission	Phase	02-Aug	20-Aug	15	
29	CW2- Presenation preperation	No	02-Aug	05-Aug	4	100
30	CW2- Presentation Submission	Yes	06-Aug	06-Aug	1	100
31	CW2 - Presentation	Yes	09-Aug	20-Aug	10	100
32	Dissertation Report Preperation	Phase	20-Jul	13-Aug	19	100
33	final dissertation report preperation	No	20-Jul	13-Aug	19	100
34	Submission of final dissertation report	Yes	13-Aug	13-Aug	1	100



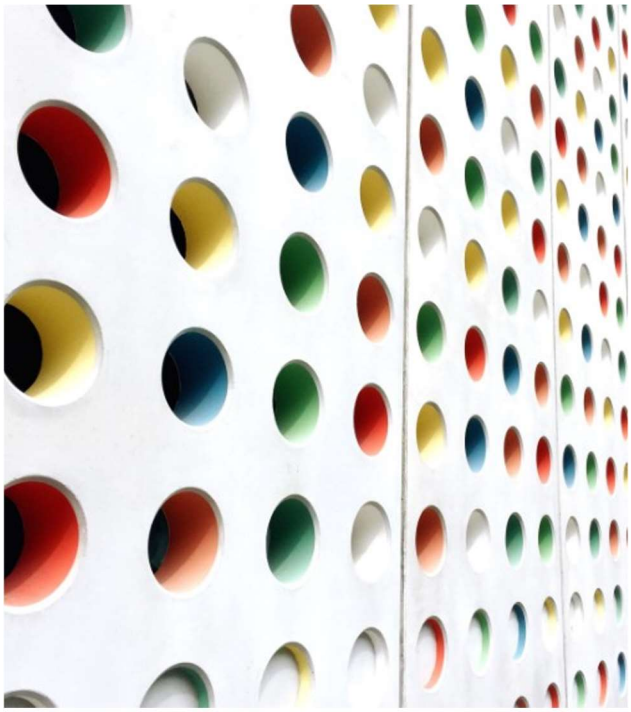
Appendix B – Interim Progress Report and Meeting Records

All the project plan and interim updates were stored in shared team folder. The folder and all its content were shared along with supervisor for any audit.
Below snapshot shows the team shared folder structure and documents in them.



Appendix D – Project Presentation

The project presentation submitted on 06 Aug 2021. The presentation slides are as follows:



Deep Learning for Soil Analysis

SUSHIL DUTT RAWAT

AUG - 2021

Project Snapshot

SOIL VISUAL-NEAR INFRARED RANGE SPECTRAL DATA EXPLORATION, DEEP LEARNING MODEL CREATION & EVALUATION, TEST PREDICTION

PREDICT 8 PROPERTIES OF SOIL USING SPECTRA DATA

9 WEEKS

Problem Statement

One of the important aspect of land surveys is to assess the soil organic and non-organic properties. Soil samples from different locations are collected using a pre-defined standard method and sent to laboratory for further analysis.

As the number of samples increase with the increase in area of land, lab testing becomes costly, and time consuming and hence alternate ways of testing is required for analysing soil properties.

Research Question

1. Can deep learning techniques predict accurate multiple soil characteristics by using vis - NIR(near infra-red) spectral data?
2. Is vis-NIR spectral data enough to predict soil characteristics from any region/continent or external information related with land coverage and land usage also required for accurate prediction?

Benefits from this research

Soil characteristics information can be helpful to farmers in deciding which type of crop is best suited and will produce maximum yield in a certain region/geography .

Government policies towards city development can be influenced based on the information about the soil characteristics . For example, fertile land is more likely to be left for agriculture compared to barren land which is more suited for extending cities and industrial developments .

State of art machine learning algorithms in soil analysis, can reduce overall analysis time and cost significantly . This will increase the number of survey frequency to provide better insight into changes in soil over the time for better management of farmlands .

Previous Work by other researchers

Data Pre-Processing

- Noise removal
- No Smoothening of Data
- Only 1 out of 8 literature reviewed applied Dimensionality reduction
- No one used Auxiliary info like elevation, Land Cover & Land Use

Prediction Algorithms

- Preferred CNN method on 1D or 2D input data for multi-output prediction
- Only 1 research with LSTM for multi-output prediction

Multi-Output predictions

- Single Target variable
- 2-5 target variable
- No research targeted all 8 output

DATASET

LUCAS (LAND USE & COVERAGE AREA FRAME SURVEY)

Lucas Dataset

Topsoil data of the 2015 LUCAS soil survey

Contains 21,859 points & 4202 feature vectors from 27 countries

Vis-NIR spectra data between 400nm-2500nm wavelength(0.5nm resolution)

Auxiliary information like elevation, Land Use & Land Class

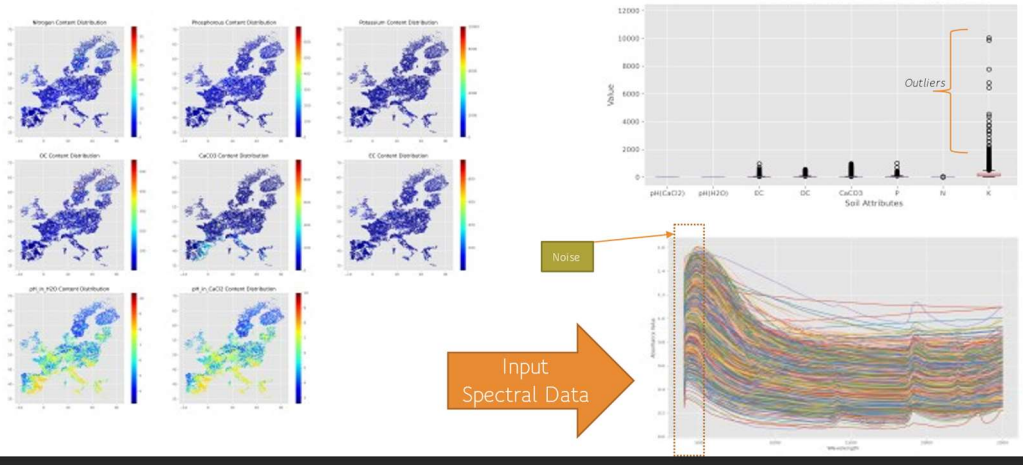
Analytical data of samples that passed the data evaluation

Field	Description	Units/Values	Limit of detection/ Measurement range
Point_ID	Unique identifier of the LUCAS survey point	8 digits number	–
Revisited_point	Is the point visited in both 2009/2012 and 2015 surveys	Yes / No	–
Coarse	Coarse elements	%	> 2000 µm
Clay	Clay content	%	0.02 – 2000 µm
Silt	Silt content	%	0.02 – 2000 µm
Sand	Sand content	%	0.02 – 2000 µm
pH($CaCl_2$)	pH measured in a $CaCl_2$ solution	–	2 – 10
pH(H_2O)	pH measured in a suspension of soil in water	–	2 – 10
EC	Electrical conductivity	mS/m	0.1
OC	Organic carbon content	g/kg	2
CaCO3	Carbonates content	g/kg	1
P	Phosphorus content	mg/kg	10
N	Total nitrogen content	g/kg	0.2
K	Extractable potassium content	mg/kg	10
Elevation	The elevation at which the soil sample was taken (in meter) (GPS_ALTITUDE)		
LC1	LandCover class		
LU1	LandUse class		
SOIL_STONES	Percentage of Stones in the sample, as a class		
NUTS_0	NUTS code for the country where the sample was taken		
NUTS_1, NUTS_2, NUTS_3	NUTS 1, 2 and 3 codes for the location where the sample was taken		
LC1_Desc	LandCover class description		
LU1_Desc	LandUse class description		

Absorbance value of Vis-NIR Spectral Wavelengths (0.5 nm resolution)

Source	SampleID	PointID	NUTS_0	SampleN	400	400.5	401			2499.5
LU_172.txt	11004_1	46442722	AT	11004	0.901924	0.907745	0.913509	----	-----	0.680794
LU_172.txt	11004_2	46442722	AT	11004	0.898363	0.904257	0.910094	----	-----	0.677908
LU_172.txt	11005_1	46522722	AT	11005	0.870422	0.876326	0.882174	----	-----	0.604204

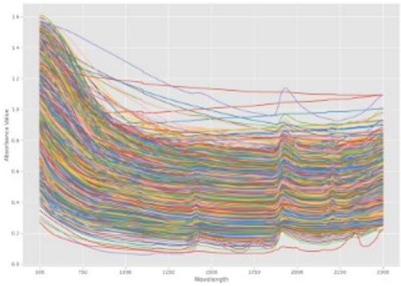
EDA - Target Variable & Input Spectral Data



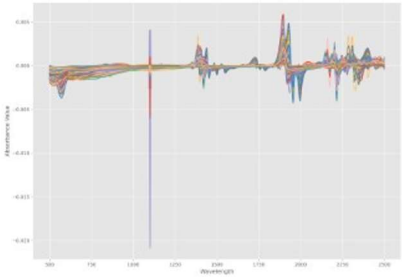
Data Pre-Processing

Noise Removal & Data Smoothening

Noise Removal : Spectral data between 400nm to 499.5nm is removed due to noise



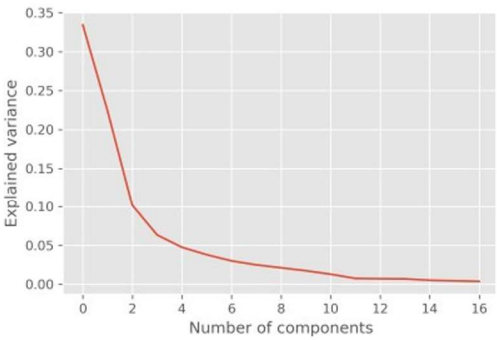
Data Smoothening - smoothing the data, to increase the precision of the data without distorting the signal tendency.

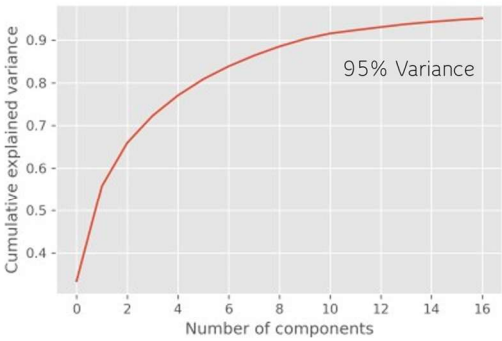


Savitzky–Golay filter

4089 Features → 17 Features

Dimensionality Reduction- PCA





Proposed Method 1- Multilayer Perceptron Network (MLP)

DIFFERENT INPUT DATA EXPERIMENTATIONS FOR MULTI - OUTPUT PREDICTION

MLP Architectures- Layers & Parameters

```
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 50)                200050
_________________________________________________________________
dense_1 (Dense)              (None, 8)                 408
=================================================================
Total params: 200,458
Trainable params: 200,458
Non-trainable params: 0
```

MLP - Input Data (without Aux information & without PCA)

```
Layer (type)                 Output Shape              Param #
=================================================================
dense_2 (Dense)              (None, 50)                204500
_________________________________________________________________
dense_3 (Dense)              (None, 8)                 408
=================================================================
Total params: 204,908
Trainable params: 204,908
Non-trainable params: 0
```

MLP_AUX - Input Data (with Aux information but without PCA)

```
Layer (type)                 Output Shape              Param #
=================================================================
dense_4 (Dense)              (None, 50)                900
_________________________________________________________________
dense_5 (Dense)              (None, 8)                 408
=================================================================
Total params: 1,308
Trainable params: 1,308
Non-trainable params: 0
```

MLP_AUX_PCA- Input Data (with Aux information & PCA)

Proposed Method 2- Deep Learning CNN

INPUT DATA RESHAPED AS 1D FOR CNN MODEL ALONG WITH AUXILIARY INFORMATION FOR MULTI -OUTPUT PREDICTION

CNN Architectures - Layers & Parameters

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 17, 128)	768
batch_normalization (BatchNo	(None, 17, 128)	512
conv1d_1 (Conv1D)	(None, 17, 128)	82048
batch_normalization_1 (Batch	(None, 17, 128)	512
max_pooling1d (MaxPooling1D)	(None, 8, 128)	0
batch_normalization_2 (Batch	(None, 8, 128)	512
dropout (Dropout)	(None, 8, 128)	0
conv1d_2 (Conv1D)	(None, 8, 64)	24640
batch_normalization_3 (Batch	(None, 8, 64)	256
dropout_1 (Dropout)	(None, 8, 64)	0
conv1d_3 (Conv1D)	(None, 8, 32)	6176
batch_normalization_4 (Batch	(None, 8, 32)	128
max_pooling1d_1 (MaxPooling1	(None, 4, 32)	0
batch_normalization_5 (Batch	(None, 4, 32)	128
flatten (Flatten)	(None, 128)	0
dense (Dense)	(None, 50)	6450
batch_normalization_6 (Batch	(None, 50)	200
dense_1 (Dense)	(None, 8)	408

Total params: 122,738
Trainable params: 121,634
Non-trainable params: 1,124

PCA_CNN_MODEL1

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 17, 128)	768
batch_normalization (BatchNo	(None, 17, 128)	512
conv1d_1 (Conv1D)	(None, 17, 64)	41024
batch_normalization_1 (Batch	(None, 17, 64)	256
dropout (Dropout)	(None, 17, 64)	0
max_pooling1d (MaxPooling1D)	(None, 8, 64)	0
batch_normalization_2 (Batch	(None, 8, 64)	256
flatten (Flatten)	(None, 512)	0
dense (Dense)	(None, 50)	25650
batch_normalization_3 (Batch	(None, 50)	200
dense_1 (Dense)	(None, 8)	408

Total params: 69,074
Trainable params: 68,462
Non-trainable params: 612

PCA_CNN_MODEL2

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 17, 64)	192
batch_normalization (BatchNo	(None, 17, 64)	256
conv1d_1 (Conv1D)	(None, 17, 32)	4128
batch_normalization_1 (Batch	(None, 17, 32)	128
dropout (Dropout)	(None, 17, 32)	0
max_pooling1d (MaxPooling1D)	(None, 8, 32)	0
batch_normalization_2 (Batch	(None, 8, 32)	128
flatten (Flatten)	(None, 256)	0
dense (Dense)	(None, 50)	12850
batch_normalization_3 (Batch	(None, 50)	200
dense_1 (Dense)	(None, 8)	408

Total params: 18,290
Trainable params: 17,934
Non-trainable params: 356

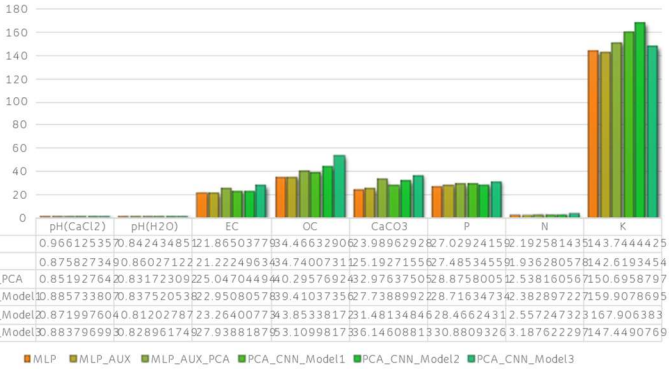
PCA_CNN_MODEL3

Input Data With Auxiliary Information and PCA Applied

Results

RMSE(ROOT MEAN SQUARED ERROR) METRICS

RMSE (Root Mean Squared Error)



	pH(CaCl2)	pH(H2O)	EC	OC	CaCO3	P	N	K
MLP	0.966125357	0.842434851	21.86503779	34.46632906	23.98962928	27.02924159	2.192581435	143.7444425
MLP_AUX	0.875827349	0.860271222	21.22249634	34.74007311	25.19271556	27.48534559	1.936280578	142.6193454
MLP_AUX_PCA	0.851927642	0.831723092	25.04704494	40.29576924	32.97637505	28.87580051	2.538160567	150.6958797
PCA_CNN_Model1	0.885733807	0.837520538	22.95080578	39.41037356	27.73889922	28.71634734	2.382897227	159.9078695
PCA_CNN_Model2	0.871997604	0.812027877	23.26400773	43.85338172	31.48134846	28.46624312	2.557247323	167.906383
PCA_CNN_Model3	0.883796993	0.828961749	27.93881879	53.10998173	36.14608813	30.88093263	3.187622297	147.4490769

- Error in prediction of pH(CaCl2 & H2O) & Nitrogen(N) is lowest compared to other properties
- Highest error is seen for Potassium(K) property (probably due to outliers in input target values!)
- Very less difference in prediction error for Phosphorous(P) across all the models
- EC prediction error is observed to be lowest for model using additional Auxiliary information & w/o PCA
- OC/CaCO3 prediction error is observed to be lowest & close for MLP models using with & without Auxiliary info but without PCA

Conclusion

- Out of all the models, MLP models performed better for many properties compared to complex CNN models.
- PCA has helped to reduce the computational complexity by reducing feature dimensionality and model performance is at par with other models without PCA.
- Simple CNN model have shown better performance than complex CNN model.
- Complex CNN models have not shown significant improvements over simple MLP models which indicate that may be CNN models are over complex for this task.
- Presence of Outliers in the target variables during training can led to high error prediction as seen in case of Potassium(K) property.

Thank You

Ethics Approval to come Here