DEEP LEARNING MODEL FOR PREDICTING SORGHUM YIELD: A CASE OF KISUMU COUNTY

SUBMITTED BY:

EDGAR M. MOSE

REG NO: 19/06924

SUPERVISOR: DR. SIMON MWENDIA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE OF THE REQUIREMENTS FOR THE AWARD OF MSC. DATA ANALYTICS IN THE FACULTY OF COMPUTING AND INFORMATION MANAGEMENT AT KCA UNIVERSITY

DECLARATION

I declare that this thesis is my original work and has not been previously published or submitted elsewhere for award of degree. I also declare that this contains no material written or published by other people except where due reference is made and author duly acknowledged.

Student name. EDGAR MATONDA MOSE……. Reg No……19/06924…..…….

Sign …  ……...…. Date……05/11/2021..……...

I do hereby confirm that I have examined the master's thesis of Edgar M. Mose and have approved it for examination.

Dissertation Supervisor:

Sign…… Date…….

Dr. Simon Mwendia

ABSTRACT

Agriculture is said to be the backbone of Kenya's economy contributing to over 20% of the country's Gross Domestic Product (GDP). More than 40% of the country's population are employed by the agricultural sector and an estimated 70% of the rural population rely on agriculture. Agricultural productivity is however dwindling owing to climate change related risks such as longer drought periods. In an effort to ensure sustainability and food security, different strategies are being implemented like climate smart agriculture which advocates for increased agricultural productivity through sustainability. Crop yield forecasting is one of the ways which can help provide useful information to policy makers and scientists to come up with sustainable agricultural strategies. It will also help farmers make informed farming decisions. Crop yield prediction is however a difficult task since many factors are considered when coming up with the ideal set of independent variables. Many studies have been conducted on predicting different crops yield using machine learning algorithms and different factors depending on the availability of data and the scope of the research. The main objective of this thesis is to come up with a deep learning model that predicts sorghum yield in Kisumu County. Deep learning is a preferred choice of machine learning algorithms because of its ability to have multiple hidden layers which increases the accuracy levels. The model will try an all-inclusive approach where all factors affecting sorghum yield production will be considered like environmental variables, agronomic, social and economic variables. Historical data obtained from the KALRO data portal will be used in this study. The Root Mean Squared Error (RMSE) and Mean Squared Error (MSE) will be used to evaluate the prediction performance of the model.

Key words: crop yield prediction, deep learning, RNN, DNN

ACKNOWLEDGEMENT

I would like to acknowledge the KCSAP project for giving me the opportunity and financial support to pursue this research. My sincere gratitude goes to my parents for their endless love and support. I would also like to thank my siblings, friends and colleagues for their support and encouragement. Lastly, I would like to acknowledge the guidance of my supervisor Dr. Mwendia whose guidance has been pivotal in coming up with this thesis.

ACRONYMS AND ABBREVIATIONS

ANN: Artificial Neural Networks

CNN: Convolutional Neural Networks

CSA: Climate Smart Agriculture

DNN: Deep Neural Networks

EABL: East Africa Breweries Limited

KCSAP: Kenya Climate Smart Agriculture Project

LASSO: Least Absolute Shrinkage and Selection Operator

LSTM: Long Short-Term Memory

MSE: Mean Squared Error

RMSE: Root Mean Squared Error

RNN: Recurrent Neural Network

SNN: Shallow Neural Network

SVM: Support Vector Machine

GLOSSARY

1. **Food security:** This is a measure of people's ability to access sufficient food that is safe and of nutritional value at all times.
2. **Crop yield:** This is a measure of the amount of crop produced per a unit area of land.
3. **Yield prediction:** This is the estimation of the amount of crop that will be produced per a unit area of land.
4. **Climate change:** This is the change in weather patterns and temperatures of a place as a result of human activities which release greenhouse gases to the atmosphere.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Agriculture plays a very vital role in the global economy and more critically in the developing countries' economies. A report by the World Bank Group (2015) points to agriculture development as one of the crucial tools to end extreme poverty since its effective ratio is higher compared to other sectors. The effectiveness of agriculture in enhancing the income of poorest countries is four times more compared to other sectors. Another report by the World Bank Group (2016) indicates that out of the working adults who are poor, 65% eked a living through agriculture.

In Kenya, Agriculture still plays a leading role towards the growth of the economy and is still a major source of employment for the populace. A Kenya Economic update report notes that agriculture has been contributing an average of 21.9% of the gross domestic product from 2013-2017 with more than half of the workforce earning their living through agriculture (World Bank Group, 2019). This sector directly influences the poverty dynamics within the country. The 17th Economic update notes that there has been a 10% decline over the years in the number of Kenyans living below the international poverty line which was estimated at individuals earning US$1.90 daily as of 2011 from a high of 46.8% to 36.1% in a span of ten years between 2005/06 and 2015/16 (World Bank Group, 2017). The country's population has also increased remarkably from 11 million in 1970 to 52.57 million in 2019. As the population rapidly increases, agricultural land on the other hand is decreasing which in turn affects food production.

Climate change is however threatening efforts of alleviating food security, poverty reduction and agriculture driven growth. Farmers have become more vulnerable to unpredictable weather patterns and drought. In order to curb climate change effects and vulnerabilities, agricultural systems need to be transformed in a manner that increases agricultural productivity while at the same time building resiliency over climate change. Some of the strategies being implemented are precision agriculture and climate smart agriculture.

Precision Agriculture as defined by Gebbers & Adamchuk, (2010) is the type of agriculture that uses a combination of different technologies like information management system, sensors and actuators, drones and enhanced machinery for agricultural production. It began in the mid-1980s with the goal of optimizing farm inputs so as to achieve maximum yields. A recent report values the global market size of precision agriculture at USD 6 billion as of 2020 with projections indicating a 13% compound annual growth up to 2028 (Grand View Research, 2021). The adoption

of precision farming can be attributed to the burgeoning Internet of Things (IoT) and big data analytics owing to easy access of information to farmers. Precision agriculture, as much as it is very effective, is also bundled with its challenges. The major limitation of precision agriculture is that it is very expensive and only the rich farmers can afford it.

Some of the factors making precision farming very expensive are like the technologies being used. The ever-evolving technology makes farmers unable to keep up as some of the software upgrades are very expensive. There are no specific standards for developing these products which in turn lead to the issue of interoperability and data fragmentation as different manufacturers are using different standards. This therefore means that data transfer and translation would require additional gateways that may prove to be a big challenge developing them. Another problem is the data generated is usually vast and may be expensive handling the storage issues and the analysis of such data as the farmers may lack the necessary skill set or will be forced to pay a professional to analyze and interpret for them. With the use of IoT comes data issues like data theft, cyber and malware attacks.

Climate-Smart Agriculture (CSA) on the other hand focuses on access and application of the agricultural data rather than optimizing on precision. It is an approach that tries to provide guidelines and effective strategies of addressing food security challenges brought by climate change. One of the major pillars of CSA is increasing the productivity and income from agriculture through sustainability. The Kenyan government in conjunction with the World bank have collaborated in the Kenya Climate Smart Agriculture Project (KCSAP) with one of their implementation components which focuses on developing agriculture related advisories like weather and market advisories (Kenya Climate Smart Agriculture Project, 2018). This thesis seeks to come up with a yield prediction model for one of the priority value chains under the KCSAP which is sorghum.

Sorghum is the only indigenous cereal species in Kenya (Duku & Groot, 2020) and comes after maize as the second common cereal crop that is grown by most of the households across the country that practice food crop farming (Onono, 2018). Sorghum is a drought tolerant crop that can grow anywhere with an altitude between the sea level and 2,500 meters above sea level. The minimum rainfall requirements for the crop are 250 mm and 10°C as the minimum temperature (Duku & Groot, 2020). The adaptation features of the crop are that it has an extensive network of roots that enable it to weather waterlogging, their leaves contain a waxy bloom which prevents loss of water and they also have the ability to suspend their growth during times of drought and

recommence back once the conditions are favorable (Muui et al., 2013). The agro-climatic zones in Kenya indicate that over 80% of the country is classified as arid and semi-arid ("Socio-economic and ecological characteristics," n.d.). These are the regions with low rainfall. This shows that with the ongoing issues of climate change and the efforts of alleviating food security, then sorghum as a crop is highly valuable.

The major zones for growing sorghum in Kenya are the Eastern and the Nyanza regions. Kitui and Kisumu Counties which are the two leading producers of sorghum contributed to almost 36% of the total yield in the country as of 2018. Kisumu produced a total of 34,666 tonnes while Kitui produced 24,327 tonnes during the long rains and 45 tonnes and 13,944 tonnes respectively during the short rains ("Sorghum production and area by counties 2018.csv," n.d.). Majority of the farmers are small scale farmers (1 to 1.5 acres) who grow sorghum alongside other crops like maize, beans and peas.

Kisumu County which is the target area in this study is a region where sorghum is widely grown and preferred. This is due to the crop's resilience and adaptation to climatic changes. Despite the decrease in area under sorghum cultivation as observed where it shrank from 11,645 ha in 2012 to 11,082 ha in 2014, the same period observed a 65% increase in production from 131,370 tonnes to 225,150 tonnes (GoK, 2015). This can be attributed to an increase in demand for the sorghum grain by industrial processors like the East African Breweries Limited (EABL) who consume an estimated 60,000 metric tonnes every year (East African Breweries Limited, 2019). Having set up a new brewery plant in Kisumu, the brewer is said to be engaging 17,000 more farmers to supply the sorghum (East African Breweries Limited, 2019).

Accurate prediction of crop yield is very vital in ensuring food production and security globally. Through accurate predictions, policy makers are able to make timely decisions on imports and exports. Farmers and other key players within the agriculture value chain stand to also benefit when it comes to making financial and management decisions as they will be acting from an informed point of view. A farmer can use the predictions to mitigate losses when hit by unfavorable conditions or capitalize on it when conditions are favorable. Successful yield prediction is however very difficult owing to the fact that there are various complex factors involved. Some of the factors crop yield depends on are like climatic conditions, water availability and quality, soil quality, pests and diseases and planning of harvesting activities (Holzman et al., 2018; Ogutu et al., 2018; Singh et al., 2016).

In looking at the factors that affect the productivity of sorghum, Okeyo et al (2020) conducted a survey and had the following findings. Land size was inversely proportional to sorghum productivity. This means that smaller farms were more productive than the larger farms owing to the fact that the small-scale farmers can easily manage smaller farms due to resource constraints. Having additional labor, especially hired labor also increased the productivity of sorghum. Different studies conducted are however conflicting on the labor issue with some finding family labor more productive than hired labor or having little significance. These issues can be imputed to different socio-economic backgrounds and geographical features that farmers have. The study also found out that the seed varieties have a significant contribution to the production of sorghum. A negative relationship was identified between the varieties of seredo and serena and the productivity of sorghum. This was attributed to other factors like untimely planting, the use of seeds that have been overstocked, poor soil fertilization due to lack of resources and loss of yields to birds. The farm gate price of sorghum indicated a positive relationship with the production of sorghum meaning that increasing the sorghum price by a unit led to an increase in the productivity. Social factors like size of the household, age, gender, total monthly income of the family and distance from the market had no significant effect on the productivity of sorghum.

The use of fertilizer on sorghum helps to considerably increase the yields (Njagi et al., 2019). However, the use of fertilizer is minimal by most of the sorghum farmers. Even the use of pesticides and herbicides to control pests and diseases is rare by farmers despite the inception of pests that have become a threat to the production of sorghum (USAID, 2013). The farmers are usually keen on having varieties that offer better yields while being tolerant to pests and diseases as the preferred qualities of sorghum.

Agricultural information is very critical in trying to increase productivity. How knowledge is transferred and agronomy is very useful in transforming agriculture from subsistence to a commercial venture. The traditional method involved having information trickle down from the Ministry of Agriculture via the use of extension officers to the farmers. These extension services have greatly declined over time due to restructuring of programs in the Ministry of Agriculture and those of county governments (Chimoita et al., 2017). Currently extension services being offered are demand driven but still limited especially to the farmers in marginalized areas (Wanyama et al., 2016). The digital era is however trying to fill that gap where farmers who have smartphones can use them to find information.

For many years the production of sorghum in Kenya has been purely for subsistence purposes. This has been gradually changing since the emergence of the malting industry which uses sorghum. The East Africa Breweries Limited (EABL) is currently leading the drive towards commercialization of sorghum production by giving market to sorghum farmers as well as attractive market prices.

Predictive modeling, which is a technique in statistics, is used to predict the future outcome of events based on historical data by use of data mining and machine learning. The models can be classified into classification, clustering, forecast, time series and outlier models. In classification data is categorized based on a query response. Clustering groups together data which shares similar attributes while forecast works with numerical data to estimate future occurrences using historical data. Time series uses time to evaluate the sequence of different data points while outlier tries to identify anomalies within the data points (NetSuite.com, 2020).

Predictive algorithms can either use machine learning or deep learning both of which are Artificial Intelligence subsets. Machine learning uses algorithms to parse data which they use for learning and then make decisions based on what they learnt. Deep learning on the other hand uses the algorithms to create an artificial neural network which imitate the human brain and can learn and make intelligent decisions on their own (Grossfeld, 2020). The deep learning algorithms are better than those of machine learning as they are non-parametric in nature hence, they have the ability to learn non-linear functions and can solve problems involving complex relationships unlike machine learning which are limited to linear problems (Pai, 2020). Another advantage with deep learning algorithms is their ability to automate feature engineering unlike machine learning algorithms which will require a domain expert to manually conduct the tasks.

The most important deep learning algorithms are; Artificial Neural Network (ANN), Deep Neural Network (DNN), Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN). The ANN is made up of three layers; the input layer that takes in the inputs, the hidden layer which computes the input and the output layer which gives the results. The DNN is a typical ANN with more layers that make them better performers compared to ANN. The RNN uses output from the previous step as an input for the next and this ability to loopback makes them ideal for making predictions for sequential and time series data. The CNN was mainly introduced for image data hence this study will focus on DNN and RNN as the candidate models for predicting sorghum yield.

1.2 Statement of the Problem

One of the important requirements towards realizing food security and poverty alleviation is having statistical data on agriculture that is authentic and can be used as evidence for making decisions. One of the ways to achieve evidence-based farming is through yield predictions. A World Bank Group report (n.d) notes that many African countries do not have the capacity to use available data for analytical studies despite the ever-increasing demand from the data users.

Accurate yield prediction is a very challenging task owing to the fact that many crop factors are put into consideration such as climatic factors, crop management factors, use of fertilizers and crop genotype and variety. Little has been done with regards to prediction of sorghum yield with available literature focusing on different factors. Zannou & Houndji (2019) created a convolutional neural network model which used sorghum images with an estimation of weight done on their ears to determine the total yield of a farm. Sridhara et al (2020) came up with different regression models to forecast sorghum yields in the Karnataka regions, India by only using weather indices. Huntington et al (2020) also used environmental variables to predict the sorghum yield biomass in the U.S.A.

There is no deep learning model that tries to accommodate diverse factors and this study aims at bridging that gap. This study proposes a deep learning model that predicts sorghum yield by factoring in climatic and environmental variables, agronomic and socio- economic variables.

1.3 Objectives

1.3.1 Main Objective

To design a deep learning model for predicting sorghum yield in Kisumu County.

1.3.2 Specific Objectives

i. To identify factors affecting sorghum yield in Kisumu County.
ii. To develop a deep learning model for predicting sorghum yield in Kisumu County.
iii. To test and validate the prediction model created.

1.4 Research questions

i. What are the ideal factors to consider when coming up with a deep learning model for the sorghum yield prediction?
ii. What is the appropriate deep learning model that can be used for sorghum yield prediction?
iii. How can the deep learning model for predicting sorghum yield be validated?

1.5 Significance of the Study

Accurate prediction of sorghum yields will be an important tool in the fight on poverty alleviation and improving food security. The analytical data generated from the sorghum yield forecasting will enable policy makers and the administration involved to come up informed agricultural decisions and strategies meant to boost agriculture which will in turn ensure food security and create employment opportunities for farmers. The farmers can also use the data to improve on their crop management practices to ensure maximum sorghum yield so as to meet the oncoming market demands. With the growing demand for the sorghum yield, more opportunities will be created for other actors within the value chain. It will enable marketers to plan well for their marketing strategies such as forward selling and storage. It will also help the industrialist when handling the logistics of the amount of yield they expect and can plan for their activities such as the required labor and maintenance. Researchers will also use the data to come up with sustainable strategies for the crop such as better yielding sorghum varieties that are drought and disease tolerant.

As commercialization of sorghum warms up, financial institutions will be able to chip in and start providing loans and other financial services to farmers. Initially many banks had perceived farmers as too risky to be offered loans given the unpredictability and risky nature of agriculture mainly due to climate change (Hong & Hanson, 2016).

In this digital era where trillions of data is generated daily, the data can only be useful if we can use it to solve our problems. This study is contributing to information technology by developing a robust model for accurately predicting sorghum yield.

1.6 Motivation of the Study

Kenya and the most developing countries in the world at large rely on agriculture as a crucial component in growing their economies. This is because agriculture is an income generating activity even to the poorest. Developments and improvements in agriculture are believed to be a solution to ending poverty and ensuring food security. A lot of agricultural data is generated on a daily basis that is mostly underutilized which if properly analyzed and put into good use can influence data driven agriculture.

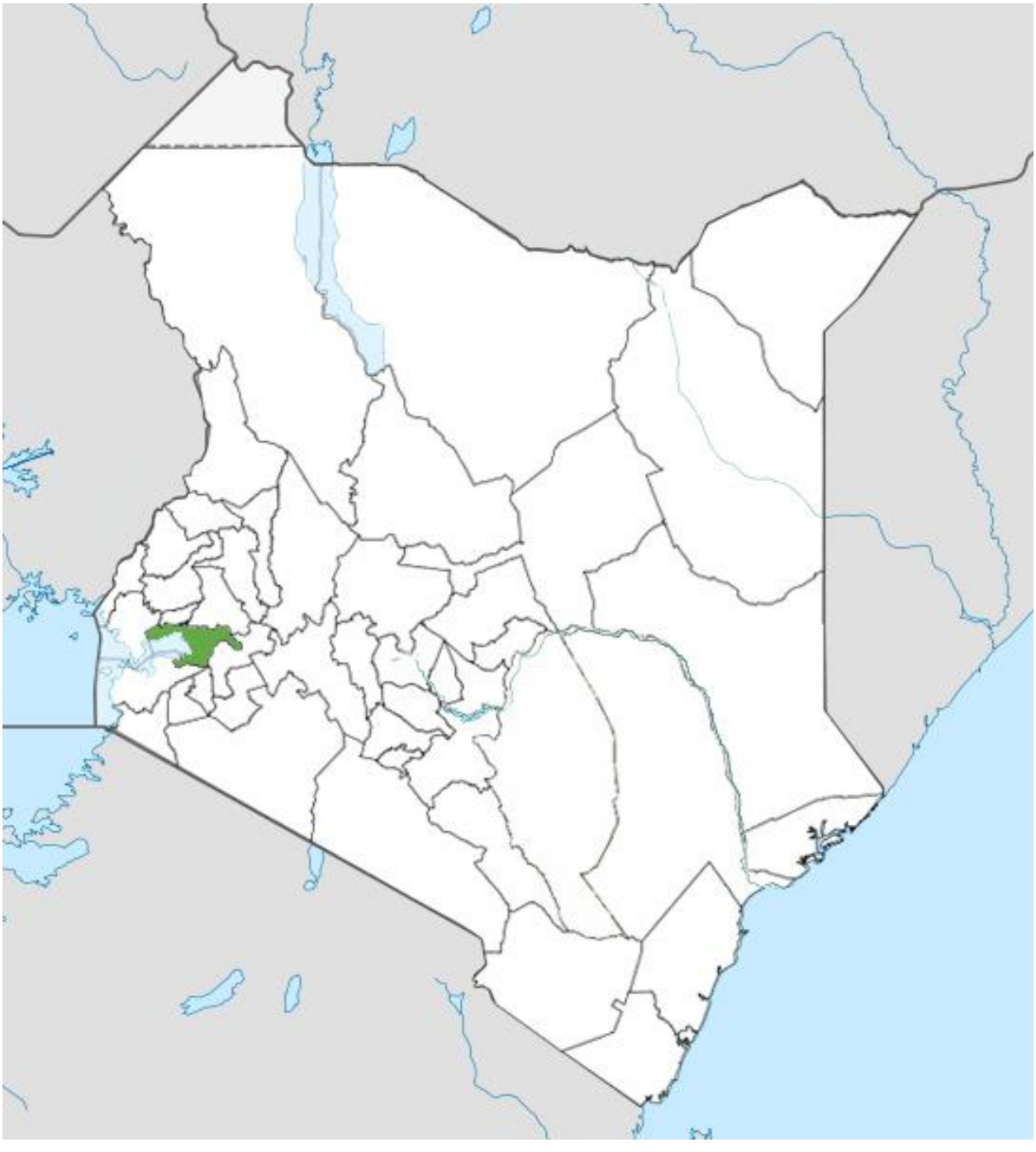
This study is therefore guided by the desire to help farmers and policy makers make informed decisions through yield predictions.

1.7 Scope of the Study

The target area for study will be Kisumu County. It is the highest sorghum producer during the long rain season in the country with a high commercial potential owing to the fact of a ready market - the EABL plant. Kenya's third largest city is located in this county. The county is bordered by Vihiga County to the North West, Nandi County to the North East, Siaya County to the West, Homa Bay County to the South and Kericho County to the East. It has a population of 1,155,574 according to the latest National census conducted in 2019. Below is a Kenyan map showing the area of study.

The technological scope of the study will only cover the deep learning algorithms of RNN and DNN alongside python programming language to come up with the yield prediction model that will help farmers in the target region gain access to agro-advisories.

FIGURE 1
The Map of Kenya Showing Area of Study



Source: (Nairobi123, 2013)

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter will provide an overview of sorghum yield, some of the factors that determine sorghum yield, an overview of machine and deep learning algorithms that can be used for yield prediction and will also analyze literary works already done by researchers on crop yield prediction. Various already developed crop yield prediction models will be explored to identify their accuracy levels, the parameters that were used and the challenges that were experienced. This will help identify what gaps need to be addressed and improved towards achieving better prediction models.

2.2 Theoretical review

2.2.1 Overview of sorghum yield

Sorghum yield is the measure of the amount of sorghum grain produced in kilograms or bags per acre of land. Sorghum yield is influenced by many factors as discussed below.

2.2.2 Factors affecting sorghum yield

a) Crop Information

This points out to information about the sorghum crop like the variety, time of maturity, weight and density of the crop (Van Klompenburg et al., 2020). Different varieties have different maturity times, special attributes and yield potential e.g., 'Mtama 1' takes 3-3.5 months to mature, its highly adaptable, tolerant to stem borers and produces an average of 15 90kg bags/acre. 'Seredo' takes 3.5 months to mature, it is tolerant to striga and yields an average of 12 90kg bags/acre. 'Serena' takes 3 months to mature, it is highly adaptable and yields an average of 12 90kg bags/acre ("Sorghum," n.d.). Varieties that take longer time to mature produce higher yields than those that have a shorter maturity time as they have a longer period for grain filling and do have an increased vegetative growth (Bandaru et al., 2006). The varieties are also influenced by the climatic conditions within which they are grown. For instance, 'Seredo' and 'Serena' thrive best in most mid-altitude areas like Kisumu.

b) Ecological factors

These are factors like climate, rainfall, temperature and soils. Sorghum relatively does well in a wide range of environmental conditions. The ideal soils for sorghum cultivation are light sandy soils and heavy clay soils with a pH of between 5.0 and 8.5. It can also survive a wide range of temperatures however; frost kills the plant and night temperatures below 12° C during flowering

period causes sterility in the plants. Below is a table summarizing some of the agro-climatic zones sorghum can grow ("Guide to farm sorghum in Kenya," n.d.).

TABLE 1
Agro-Climatic Zones

Agro-climatic zone	**Altitude (meters above sea level)**	**Annual rainfall (mm)**
Semi-Arid lowlands	250 - 1500 m	250 - 500 mm
Mid-semi-arid highlands	1750 - 2000 m	300 - 800 mm
Humid coastal lowlands	0 – 250 m	400 – 800 mm
Moist-mid altitude	1150 – 1750 m	500 – above mm
Cold semi-arid highlands	1750 – 2000 m	300– 800 mm

c) Field management

This is the making and implementation of decisions around the field so as to achieve optimum yields. This entails activities like fertilizer application, irrigation and pest and disease control. Fertilizer application is recommended for enhancing sorghum yield especially organic foliar feeds or use of compost manure. Intercropping sorghum with a legume or rotational cropping with a legume is also encouraged for bettering yields (KALRO, n.d.). The need for irrigation can be determined by the amount of rainfall the area receives and the water retention capabilities of the soil. Temperature and root depth can also be considered. Some of the visible signs a plant can show is the withering of leaves during the morning hours and cracked soil surface ("Irrigation in sorghum," n.d.). Sorghum is highly disease tolerant and application of agro-chemicals is rare.

d) Socio-economic factors

These are factors like age, size of the household, gender, farm size, education level, land ownership, labor and access to credit. The age of farmers is an important aspect in agriculture as it helps to identify the experience a farmer has in farming. The average age for farmers is over 40 years for most of the studies done (Mmbando & Baiyegunhi, 2016; Ogeto et al., 2013; Suvedi et al., 2017). This suggests that farming is mostly done by the elderly which confirms the claims that while the elderly focus on farming, the youth focus on off-farm activities like business and formal jobs (Demissie, 2013). Ogeto et al., (2013) in their research identified that there were more women (54.6%) in sorghum farming than males (45.4%). This confirms the notion that women focus more on subsistence farming while men focus more on commercial farming (Curran & Cook, 2009). A large household size is an indication of farm labor that is readily available. This has a positive

influence on sorghum yield. Okeyo et al., (2020) in their study found out that farm size and training on good sorghum variety had a positive effect on sorghum adoption. Formal education had a negative effect as individuals with formal education prefer to work in off-farm jobs. Land ownership was identified to also have significant negative effect as the owners are less likely to participate in the production of sorghum unlike those who rent land with the intention of profit maximization. The size of the farm will dictate what a farmer will plant. Ogeto et al., (2013) found out that farmers with larger pieces of land tend to participate in the production of sorghum unlike those with smaller farms. Access to credit and availability of market for the sorghum yield also has a positive effect on sorghum production (Chimoita et al., 2017).

2.3 Machine learning algorithms

These are the type of algorithms that allow machines to identify and learn different relationships within a dataset and use the knowledge and patterns learnt to make predictions. These algorithms can be broadly classified into Supervised and Unsupervised learning algorithms. Supervised learning algorithms use labelled datasets to train a model while Unsupervised learning algorithms learn from unlabeled datasets to train their models. Supervised learning algorithms which this study will use is also categorized into classification and regression models. Regression models identify patterns and makes predictions for continuous data while classification is used when the output consists of categorized values. Since this study will be working on continuous data the choice of models will be regression models. Examples of regression models are as below:

i. Linear regression – This algorithm is used to predict the value of the dependent variable by using identified independent variables through identifying a linear relationship between the dependent and the independent variables. It is a simple technique to implement but suffers from multicollinearity and suffers from outliers.
ii. Decision tree – This is an algorithm that can be used for both classification and regression. The decision trees can easily capture non-linear relationships between the selected features and the target variable. The advantage for this model is that it requires little preprocessing of the data but the disadvantages is that it suffers from overfitting and a slight change of data causes a big disruption in the whole tree (Sharma, 2021).
iii. Support Vector Regression – This algorithm uses a hyperplane to separate data. If the separation is not successful then a kernel trick is employed which increases the dimension to levels a hyperplane can be able to split the data points (Sharma, 2021). The advantages of this algorithm are that it is robust to outliers and has high prediction accuracy. The disadvantages of this model are that they do not work well with large datasets or data with a lot of noise.

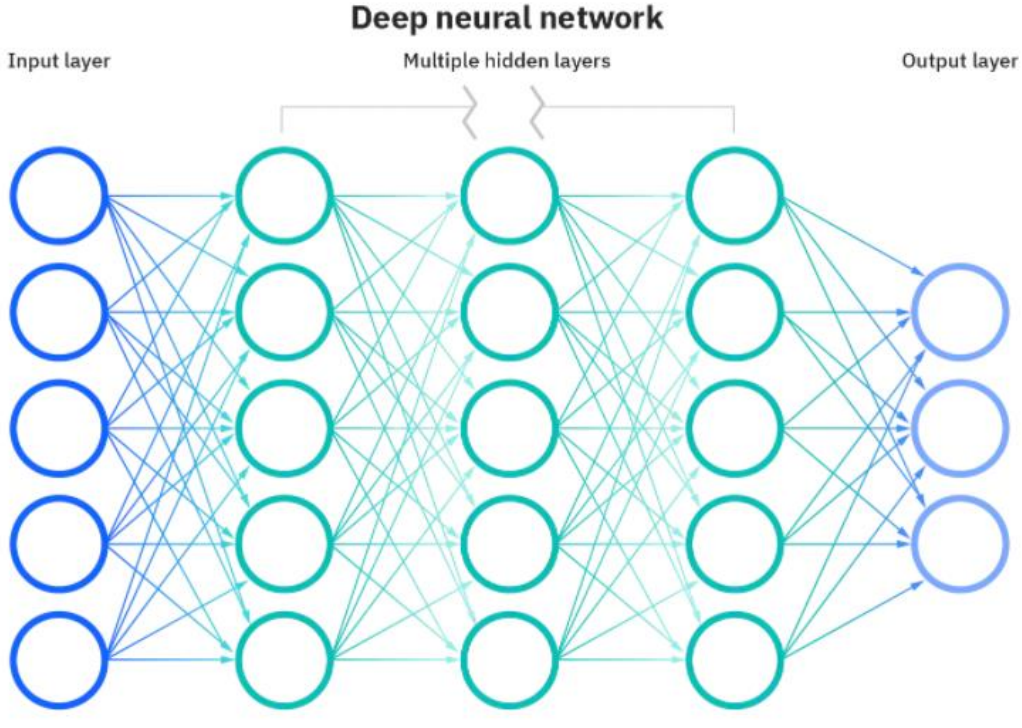
iv. Lasso regression – These algorithm works by identifying and applying a constraint on model attributes such that those that don't meet the threshold are shrank to zero. It is good for feature selection and prevents the model from overfitting. However, the features selected maybe biased and lasso will only select one feature from a group of correlated features.

2.4 Deep learning algorithms

Deep learning is a subset of Artificial Intelligence and machine learning which consists of a collection of algorithms which use multi-layered artificial neural networks to process and compute large amounts of data to create patterns which can be used for decision making. They mimic the structure and how the human brain works.

A deep neural network is made up of more than three layers. The input layer, the hidden layers and the output layer. The input layer passes information to the hidden layer by firing an activation function. The hidden layers fine tune the input weights in the neural network through activation of the weighing function until a value that meets a particular threshold is found which is then sent to the output layer. The Figure 2 below illustrates the deep neural network.

FIGURE 2
Deep Neural Network



Source: ("AI vs. machine learning vs. deep learning vs. neural networks: What's the difference?," 2020)

The most commonly used technique by the deep neural network is the feed-forward technique which means that their direction of flow is from input to output. However, a model can be trained via backpropagation which is movement in the opposite direction from output to input. The advantage of backpropagation is that it allows one to calculate and identify errors attributed to a particular neuron and make the necessary adjustments so as to suitably fit the algorithm ("AI vs. machine learning vs. deep learning vs. neural networks: What's the difference?," 2020).

Deep learning algorithms can be broadly classified into two; supervised deep learning algorithms and unsupervised deep learning algorithms.

Supervised deep learning algorithms.

This is the use of a labelled dataset to train an algorithm into predicting the outcome or classifying a dataset. They are mainly used to handle tasks related to classification and regression ("Supervised vs. unsupervised learning: What's the difference?," 2021). Some of the examples of supervised deep learning algorithms are:

i. Artificial Neural Network (ANN) – This is a modeling technique which tries to replicate the nervous system of a human being. It tries to learn through data presented to it and finds hidden relationships between independent and dependent variables. The ANN model is made up of node layers which contain the input layer, one or more hidden layers and the output layer which have already been illustrated above. The model first goes through a training phase where the network makes comparison between the actual output generated and the desired output. The difference between the two outcomes is then adjusted via backpropagation technique until the lowest possible error is realized ("Artificial neural network (ANN)," n.d.). ANNs require large datasets for training so as to increase their accuracy.

ii. Convolutional Neural Network (CNN) – The CNN can be compared to a fully connected network. They consist of three types of layers; the convolutional layer, the pooling layer and the fully connected layer. The convolutional layer is made up of feature maps and filters. The filters act as the neurons of the layer and have weighted inputs which generate the output value (Brownlee, 2016). The output of a filter can be regarded as a feature map. Pooling layers are used to reduce overfitting by downsampling the preceding layer's feature map, generalizing feature representations, and downsampling the prior layer's feature map (Brownlee, 2019). For predictions, the fully-connected layers are typically employed at the network's end. The general structure for the CNN model starts with one or more convolutional networks which is followed by a pooling layer. The process is then repeated severally until the application of fully

connected layers on the end. This algorithm is ideal for computer vision, image and pattern recognition applications.

iii. Deep Neural Network (DNN) - The DNN is an upgrade of the ANN network since it has more hidden layers which are almost wholly connected and learn using lesser parameters. The many hidden layers make them perform much better compared to the ANN models.

iv. Recurrent Neural Network (RNN) – This algorithm treats output from the previous step as an input for the next step. The process of passing back information into the network by looping from output to input makes the model to easily remember past information and use it for future predictions (Karagiannakos, 2020). This ability makes them an ideal algorithm for working with time series and sequential data.

Unsupervised deep learning algorithms.

This is where the algorithms analyze and group unlabeled datasets. They do not require human intervention to identify hidden patterns in data. They are mainly used for tasks that require association, clustering and dimensionality reduction. Some of the examples of unsupervised deep learning algorithms are:

i. Autoencoders – These algorithms to reconstruct data by making sure the output is equal to the input. They consist of an encoder which encodes the input they receive into a lower dimensional latent space and a decoder that decodes it back to the original input (Karagiannakos, 2020). They are mainly used for compression and dimensionality reduction.

ii. Restricted Boltzmann machine (RBM) – This a stochastic neural network with the capabilities of using their inputs to learn their probability distribution. This network has only two layers; the input and the hidden layers. They work like autoencoders with the only difference being that they only use a single network. The feed forward involves producing a representation from a given input and a backpropagation involves using the representation to reconstruct the original input (Karagiannakos, 2020).

iii. Deep Belief Networks (DBN) – A DBN is formed by stacking multiple RBMs. They resemble fully connected networks with the difference being how they are trained since the DBNs train the layers in pairs (Karagiannakos, 2020).

The deeper models aside from their high accuracy do experience challenges too. They are more difficult to train and necessitate more sophisticated hardware and optimization approaches (Goodfellow et al., 2016). They also experience problems with loss functions and the vanishing gradient problem. The Deep neural networks' loss function is highly dimensional and non-convex,

making optimization more challenging due to the presence of multiple local optima and saddle points (Goodfellow et al., 2016). The vanishing gradient problem is a problem in which the gradients of the loss function shrink exponentially making it difficult to train. This happens with the addition of more layers to the neural network using bounded activation functions. This problem can be fixed by using activation functions which are not bounded like ReLU (Wang, 2019).

From machine and deep learning algorithms it is evident that the deep learning algorithms are superior as they are flexible in terms of fine tuning and adjusting of the hyperparameters to achieve better results. Therefore, for this study, the best models identified are the RNN and the DNN which will be developed and compared to find the best model for this study.

2.5 Empirical Review

Many yields prediction models and applications have been developed over time. They have been majorly classified as Statistical models and Crop Simulation models (Yadav et al., 2019). These methods however work with linear data hence are less accurate when dealing with the non-linear effects of crop yield factors. The introduction of Artificial Intelligence models and applications is proving to be more efficient and accurate in making predictions compared to the traditional models. Artificial Intelligence has subsets like Machine Learning and Deep Learning techniques that are used to provide solutions to real world issues. In the agricultural frameworks, the algorithms go through a learning process where the dataset is trained to perform a specific task with the required output represented using past experiences. Once done with the training process, the models are allowed to make presumptions on the test data (Elavarasan & Vincent, 2020).

Some of the previous works done in relation to crop yield prediction using deep learning algorithms are like Khaki & Wang (2019) who used deep neural networks to predict hybrid corn yield of different locations in the USA using environment and genotype data. The training data had three datasets; the crop genotype, environmental data which consisted of weather and soil data and the yield performance data. Three other models were implemented for comparison. These are; Regression tree, least absolute shrinkage and selection operator (Lasso) and Shallow neural network (SNN). The DNN model outdid the rest of the models with the least RMSE (root mean square error) of 11%. The major drawback of the model was the black box issue but a feature selection was performed using backpropagation to try and diminish the effects which revealed that crop yield was greatly affected by environmental factors compared to genotype. Shook et al., 2018 in their research also used genotype and weather variables to predict soybean crop yield. A LSTM

and temporal attention model were used so as to capture the temporal variability of disparate weather variables.

Khaki et al (2020) came up with a hybrid model which was made up of a convolutional neural network (CNN) and a recurrent neural network (RNN) model to predict corn and soybean yield in the corn belts of the USA. They used environmental data, management practices data and yield performance data of the crops. No genotype data was used. The model was implemented alongside Random Forest (RF), LASSO and deep fully connected convolutional neural networks (DFNN). With a RMSE of 9% the model outdid the rest of the models. Just like the DNN model, this hybrid model also had the black box property limitation where a feature selection was done through back propagation to estimate the effects of individual factors against the crop yield. A comparison evaluating the importance of the main factors shows that soil and weather factors produced a higher prediction accuracy than management practices.

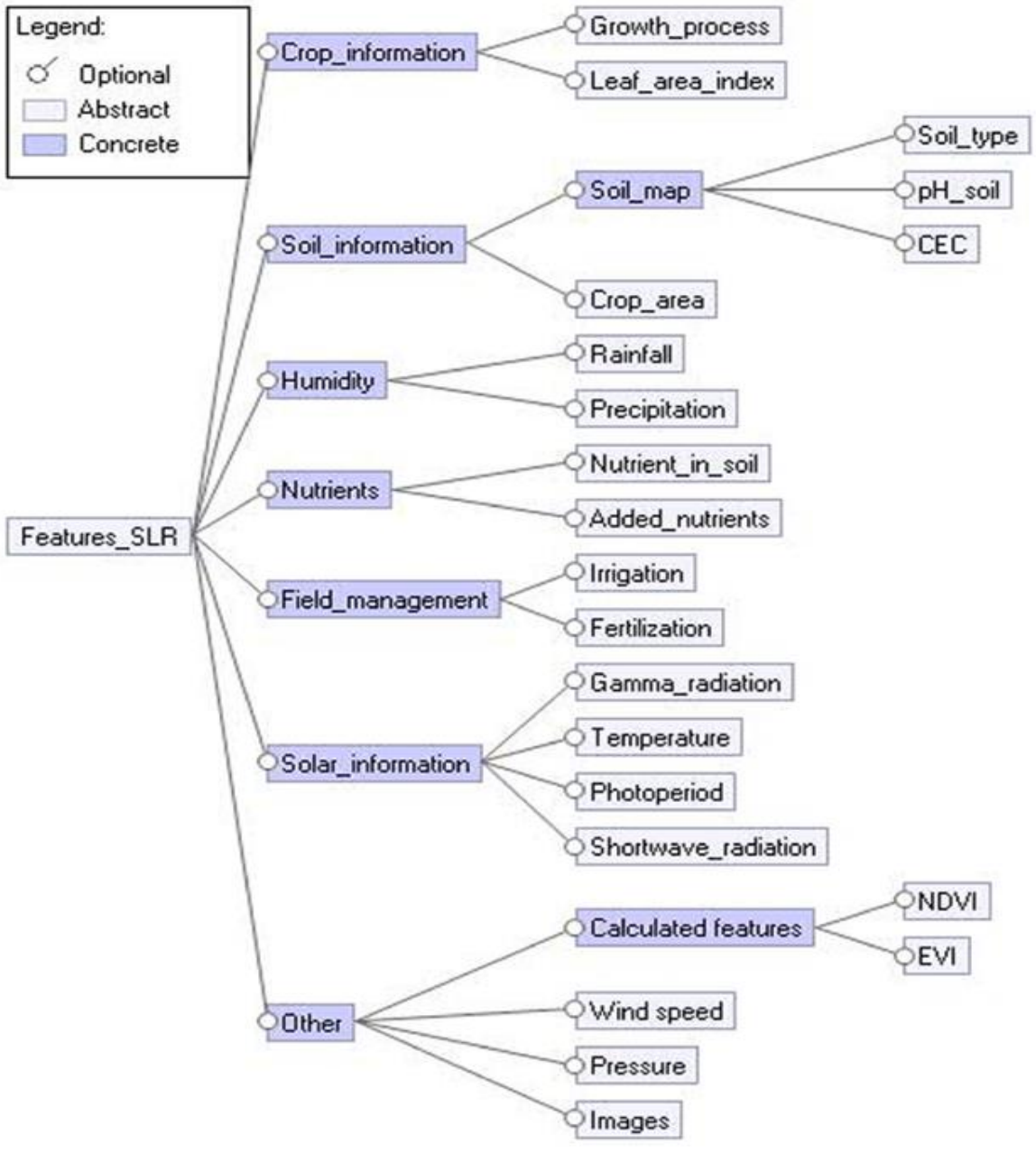
In using a hybrid approach to predict crop yield, Agarwal & Tarar (2021) used LSTM and RNN as the deep learning algorithms alongside Support Vector Machine which is a machine learning algorithm. The data used was climatic and soil data with crops being tested like wheat, maize, rice and peas. In analyzing the performance, ANN was combined with RF to achieve an accuracy of 93% while the hybrid of LSTM, RNN and SVM achieved a 97% accuracy. It was concluded that deep learning algorithms enhance the accuracy of a model.

Elavarasan & Vincent (2020) came up with a model that predicted rice yield in Southern India using a Deep Recurrent Q-Network model. This model consisted of a RNN model over a Q-learning reinforcement algorithm. A dataset of 35 years was used which had climatic and soil factors. A K-fold cross validation was done to ensure every important information is factored in which led to an accuracy of 93.7% outperforming other models tested alongside it like LSTM, ANN and RF.

In all the models discussed above we realize that the dominating gap is the different factors used for yield predictions. In as much as there could be multiple factors affecting the crop yield, the models have been limited to a couple of factors. Van Klompenburg et al (2020) who conducted a systematic literature review on different studies done on crop yield prediction using machine learning also note that each paper tries to predict yield but uses a different variety of features. This could be influenced by the availability of data and the scope of the research. There is no model that tries to use all the factors identified to be affecting the crop yield. Through their study, Van

Klompenburg et al (2020) came up with a feature map to highlight some of the significant features mostly used and their sub features as shown in Figure 3 below.

FIGURE 3
Feature Diagram

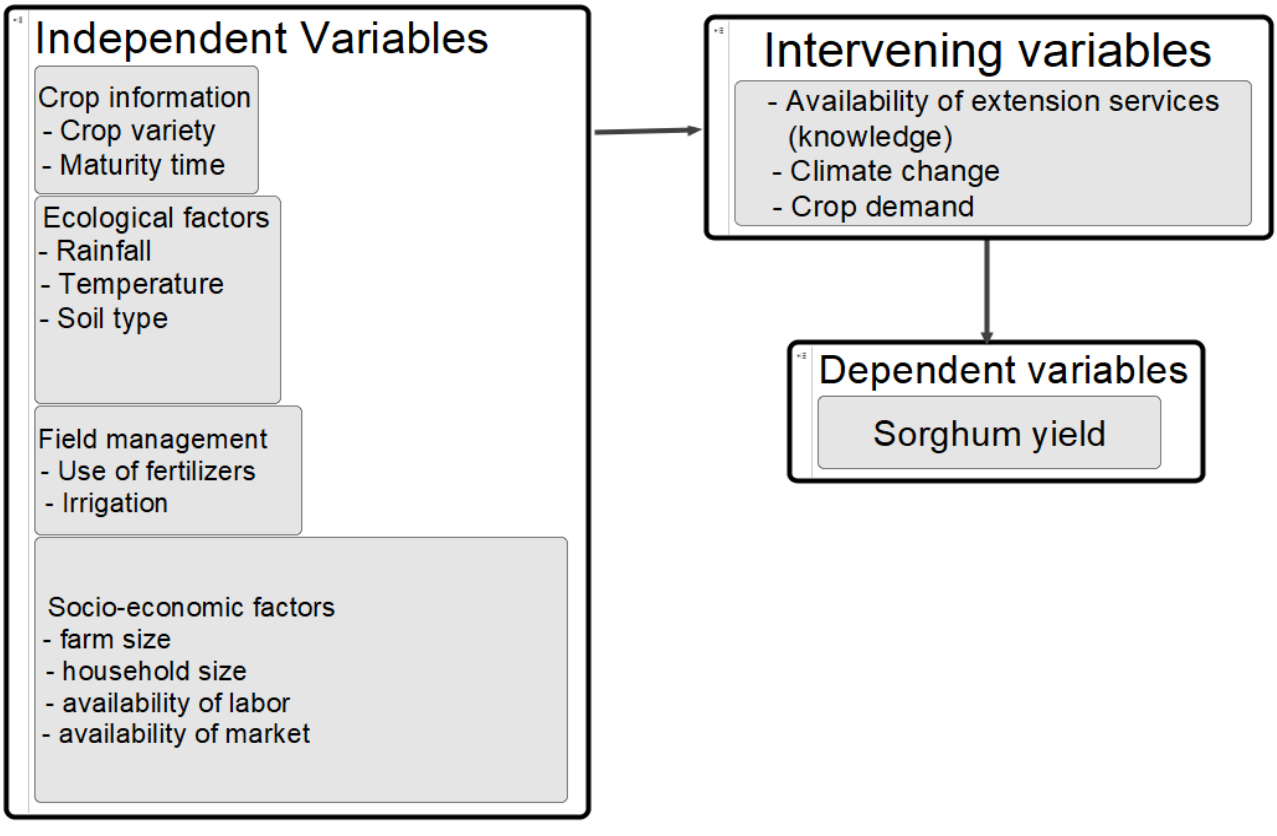


Source: (Van Klompenburg et al., 2020)

2.6 Conceptual Framework

As guided by the literature review, the following conceptual framework was developed. The Figure 4 below visually represents the relationship between that of the independent variables, the intervening variables and the dependent variables.

FIGURE 4
Conceptual Framework



Source: Author (2021)

2.7 Operationalization of Variables

TABLE 2
Operationalization of Variables

Variables	**Sub-variables**	**Indicators**	**Values**
Factors influencing sorghum yield	Crop information	Crop variety	Serena, Seredo, Mtama1
		Maturity time	3, 3.5
	Climatic factors	Rainfall	Integer
		Temperature	Integer
		Soil type	Sandy, Loamy
	Field management	Use of fertilizers	Yes, No
		Irrigation	Yes, No
	Socio-economic factors	Farm size	Integer
		Household size	Integer
		Availability of labor	Yes, No
		Availability of market	Yes, No
Sorghum yield		Sorghum yield	Integer

Source: Author (2021)

2.8 Summary

This chapter mainly reviewed some of the literary works done on crop yield prediction. Specific focus was towards the objectives of this research which is identifying the factors to consider for sorghum yield prediction and the ideal deep learning model to be used for predicting the sorghum yield. The next chapter will talk about the methodology that will be employed in this research.

CHAPTER THREE

METHODOLOGY

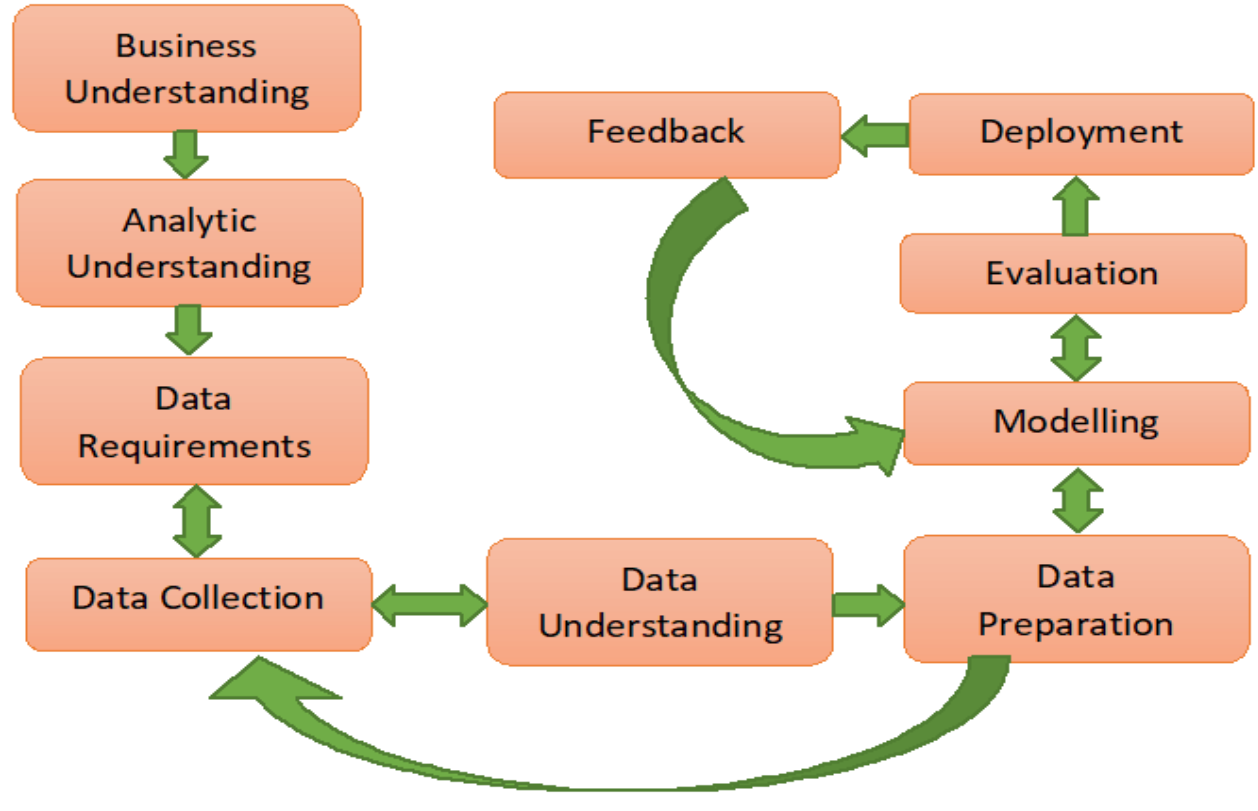
3.1 Introduction

This chapter discusses the research design that has been adopted so as to achieve the objectives of this study. It also highlights the target population, how sampling will be done, the data collection process and the instruments used, how data will be processed and analyzed.

3.2 Research design

This project will use the data science methodology. This is a simplified methodology that helps data scientists to find solutions to their business problems by use of involved data. Without a proper methodology, even with the advancement of technology and access to data, the fundamental questions of problem identification and understanding of how properly data should be analyzed would lead to poor decision making (Patel, 2019). The choice for this methodology is supported by its ability to provide a step wise simple problem-solving approach. The methodology consists of ten steps that can be summarized into three main questions.

FIGURE 5
Data Science Methodology



Source: ("Data science methodology and approach," 2019)

The first two questions are concerned with problem identification and the approach that will be used in solving the problem. The next four questions are concerned with the data identification and manipulation while the last four are concerned with how a solution will be derived.

i. **Business understanding**: This a very crucial step towards problem solving as well understood problems lead to better and quality solutions at the end of the process. It is through this step that business requirements are elicited. In this case we are trying to come up with a model of predicting sorghum yield.
ii. **Analytic understanding:** The analytic approach being used in this project is the predictive analytics where we are trying to predict sorghum yield. Predictive analytics is a technique of trying to estimate the outcome of future events by use of historical and current data.
iii. **Data requirements**: The analytical approach identified dictates what data is required for the project. The variables required in the yield prediction will include historic climatic data, historic yield performance data, optimal climatic data for optimal yields, agronomic practices done by farmers, sorghum varieties, farmer household data like; size of land, source of labor and market availability.
iv. **Data collection:** Historical sorghum data from the KALRO data portal will be used in coming up with the model.
v. **Data understanding:** Trying to understand whether the data collected represents the problem about to be solved. This involves understanding the data types and their respective attributes.
vi. **Data preparation:** Involves data preprocessing like data cleaning. More about data preparation is discussed below.
vii. **Modelling:** Predictive data analytics will use the deep learning algorithms. The choice of algorithms will be Recurrent Neural Networks (RNN) and the Deep Neural Networks (DNN) to predict the sorghum yields.
viii. **Evaluation:** The Root Mean Squared Error (RMSE) and Mean Squared Error (MSE) will be used in trying to evaluate the prediction performance of the models.
ix. **Deployment:** After evaluation of the model and being sure it is working the model can then be deployed and put out to real test.
x. **Feedback:** This is the feedback provided by the users once the model is in play. It provides room to refine the model and assess its impact and performance.

3.3 Target population

This study will utilize historical agricultural data, specifically sorghum data from Kisumu County. The data will consist of records of daily climatic data, seasonal yields between 2016-2020, sorghum crop information, field management and the social economic factors affecting sorghum yield. This data will be retrieved from the KALRO data portal.

3.4 Sampling and Sampling procedure

This research intends to sample historical sorghum data for a period of five years between 2016-2020 which consists of 45626 rows. Deep learning models are known to be non-parametric with high flexibility which makes them need more training data so as to increase their accuracy. There is no known formula to determine how much data is enough data with most models depending on the availability of data and scope of their models. Nasir and Sassani, (2021) however advise that for a deep learning model the number of samples should be tenfold the number of parameters. The granularity of a five-year data would suffice for making yield prediction.

3.5 Data collection procedure

The process of data collection will involve logging into the KALRO data portal and downloading the identified sorghum datafiles.

3.6 Data processing and analysis

The processing of data will involve data cleaning which is the removal or replacement of inconsistent data. The first step will involve merging of the datasets. The data will then be deduplicated, handling of missing data, removal of incorrect and irrelevant data and fixing of structural errors like inconsistent use of capitalization and typos. The data will then be standardized and normalized for easy analysis.

Once the data is cleaned, deep learning predictive models will be used to predict the sorghum yields. All these processes will be done by use of the python programming language on the Jupyter notebook environment.

3.7 Model validation

Model validation involves a set of processes aimed at verifying whether the model is able to achieve the intended tasks. This study will validate the model using the K-Fold Cross-Validation technique. This technique splits the data into k folds and uses the k-1 data for training and the remainder for testing. The advantage of this technique is its ability to minimize bias in sampling with all the observations being used for both training and validation and can only be used once. This technique enables the models to easily learn the distribution of the dataset.

CHAPTER FOUR

DATA ANALYSIS, FINDINGS AND DISCUSSION

4.1 Introduction

This chapter presents the analysis of the collected data and discusses findings from the study. Secondary data of sorghum yield was collected and analyzed using python libraries with the guidance of the objectives set to be achieved. Presentation of the data has been done using descriptive tools for ease of interpretation.

4.2 Descriptive Statistics

This study focuses on sorghum yield data for Kisumu County that was collected from 2016-2020. The data consists of six sorghum growing sub-counties which are Kisumu West, Kisumu East, Muhoroni, Nyakach, Nyando and Seme. Every year consists of two planting seasons which are based on the long rain season and the short rain season. The data is made up of different factors which contribute to sorghum yield. Below is a snippet of the dataset.

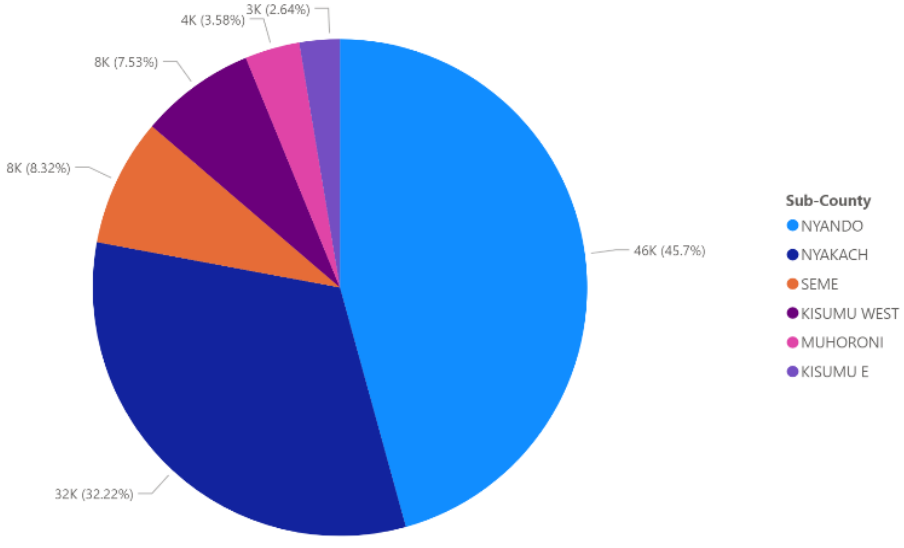
FIGURE 6

Sample Dataset Snippet

	Sub-Coun	Year	Area(Ha)	Ppt (mm)	Max Temp	Min Temp	Soil Type	Irrigation	Fertilizer a	Seredo	Serena	Indigenou	Andiwo	Red sorgh	Kari mtam	Kari Mtam	Maturity	Yield (T)
2	KISUMU W	2016	700	853.5235	28.03156	18.2591	Red-loamy	0	0	1	1	0	0	0	0	0	100	378
3	SEME	2016	1150	744.5655	28.39902	19.41328	Sandy-clay	0	0	1	1	1	0	0	0	0	112	621
4	KISUMU E	2016	365	833.66	28.33527	18.62799	Black-cott	0	0	1	1	1	1	0	0	0	98	197
5	NYANDO	2016	6350	679.3775	28.59265	19.26832	Black-cott	0	0	1	1	0	0	1	0	0	108	3429
6	NYAKACH	2016	4470	784.288	28.58713	19.59049	Red-loamy	0	0	1	1	0	1	0	1	0	115	2414
7	MUHORO	2016	500	844.5963	28.40027	18.67557	Black-cott	0	0	1	1	1	0	0	0	0	118	270
8	KISUMU W	2017	836	741.747	28.36267	17.85261	Red-loamy	0	0	1	1	1	0	0	0	1	101	827
9	SEME	2017	1373	763.6317	28.42293	18.37838	Sandy-clay	0	0	1	1	0	0	0	0	0	106	1359
10	KISUMU E	2017	436	740.8228	28.7161	18.50294	Black-cott	0	0	1	1	1	0	0	0	0	114	432
11	NYANDO	2017	7583	688.3337	28.33838	18.18714	Black-cott	0	0	1	1	0	0	0	1	0	112	7507
12	NYAKACH	2017	5338	695.7901	28.56425	18.46647	Red-loamy	0	0	1	1	1	0	0	0	1	106	5285
13	MUHORO	2017	597	647.5325	29.06787	19.1927	Black-cott	0	0	1	1	1	0	0	0	0	121	591

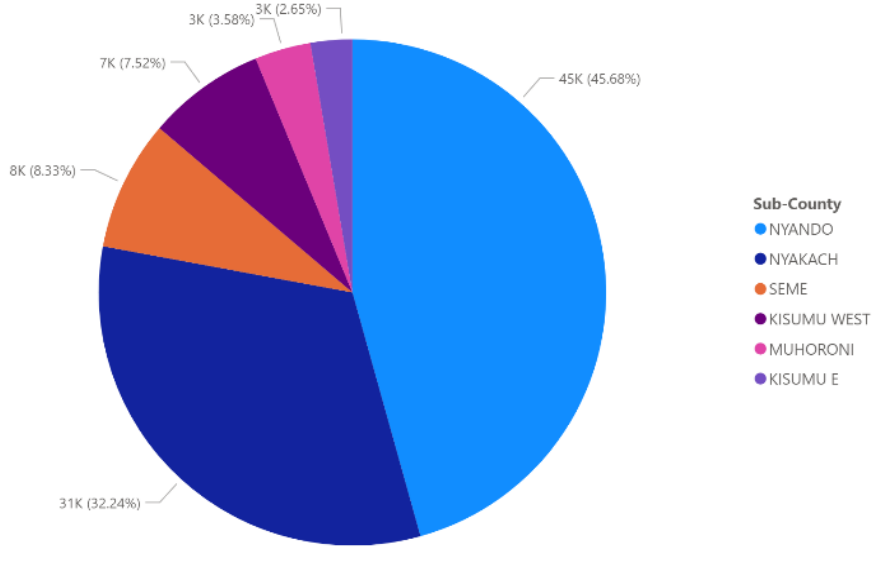
A descriptive analysis was done on the dataset of the different factors that contribute to sorghum yield. Below are images describing the findings of the data collected.

FIGURE 7
Proportion of area by Sub-County



Source: Author (2021)

FIGURE 8
Proportion of Yield by Sub-County

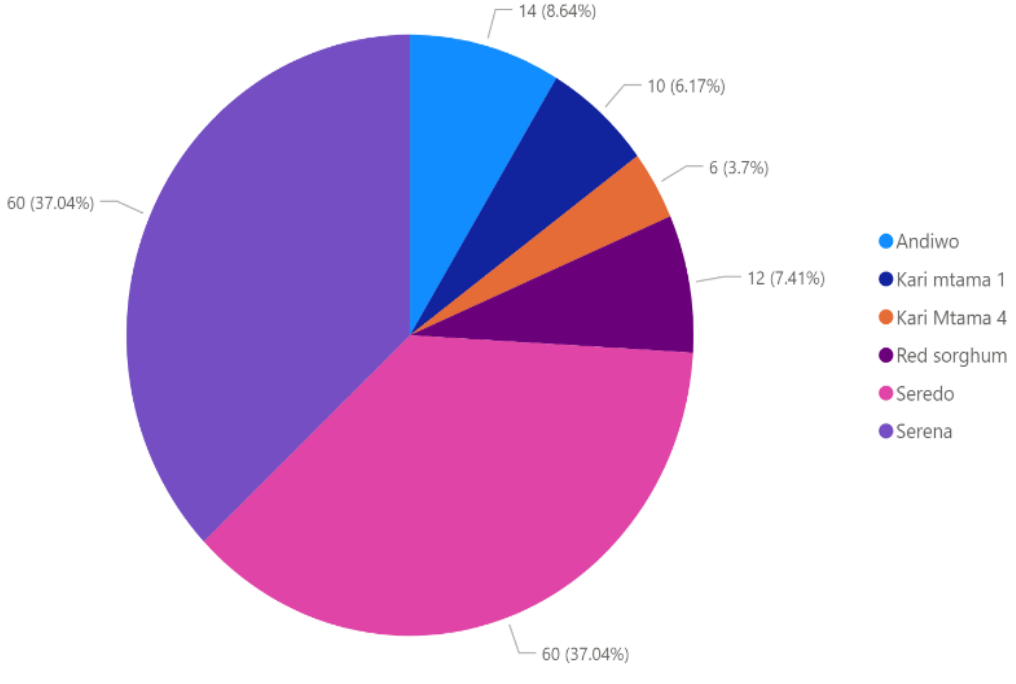


Source: Author (2021)

Figure 7 above shows the proportion of sorghum growing areas by each Sub-County. Nyando Sub County provides almost half of the total area of Kisumu County that grows sorghum at 45.7% while the least growing Sub County which is Kisumu East only contributes to 2.64% of the land. Figure 8 above which shows the proportion of yield by sub county also notes that Nyando leads with 45.68% while Kisumu East comes last with a contribution of 2.65%. The two figures imply that land area is directly proportional to yield in that the higher the land size the more the yield.

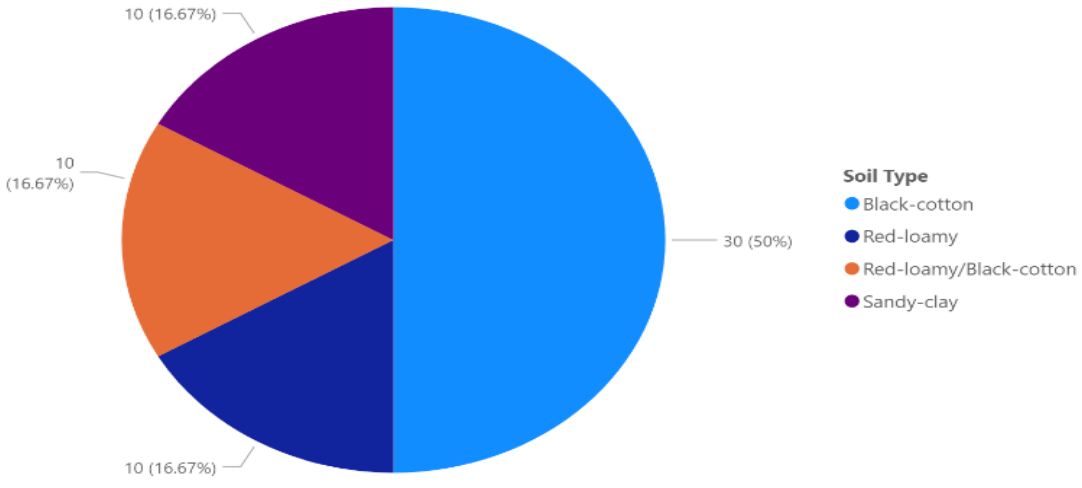
Figure 9 below shows the quantitative representation of the mostly planted sorghum varieties. The most common sorghum varieties being planted are the Seredo variety and Serena varieties which tie at 37.04% each while the least planted is the Kari mtama 4 which is at 3.7%. The major soil types identified within the county were the black-cotton soil, red-loamy soil, sandy-clay soil and a mixture of red-loamy and black-cotton. The black-cotton soil was the most dominant at 50% with the rest of the soils taking up 16.67% each.

FIGURE 9
Proportion of Sorghum Varieties Mostly Planted



Source: Author (2021)

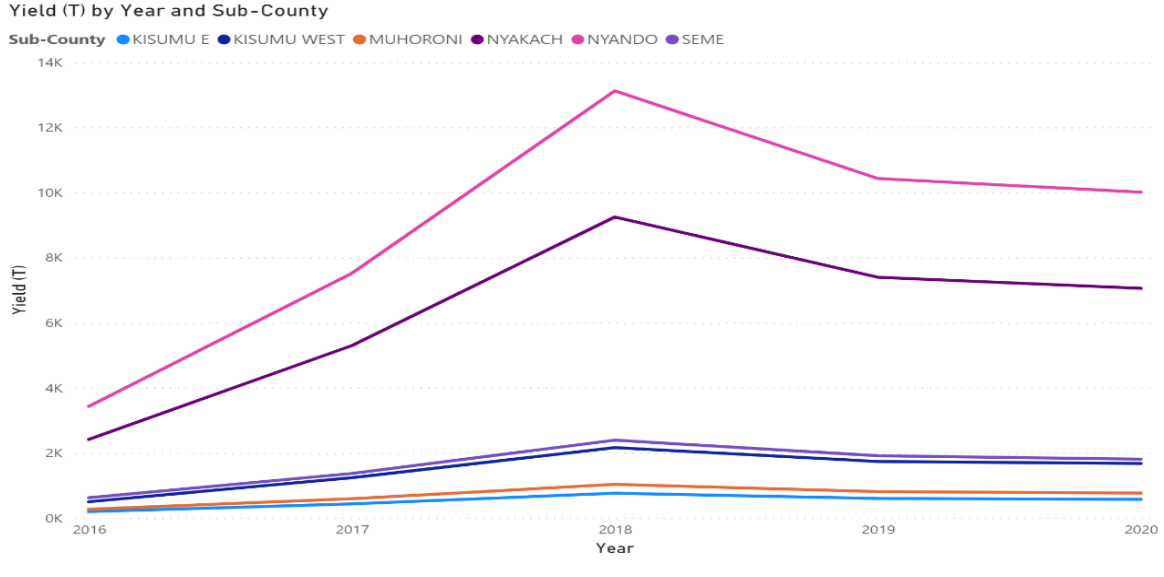
FIGURE 10
Proportion of Soil Types



Source: Author (2021)

In looking at the yield trends per year for each Sub County, line graphs were plotted as shown in Figure 11 below. Nyando exhibits the highest yields over the years with the highest yield peaking at 2018. The same applies to all the other sub counties who have had their best yield in 2018.

FIGURE 11
Line Graph Showing Yield by Year and Sub-County

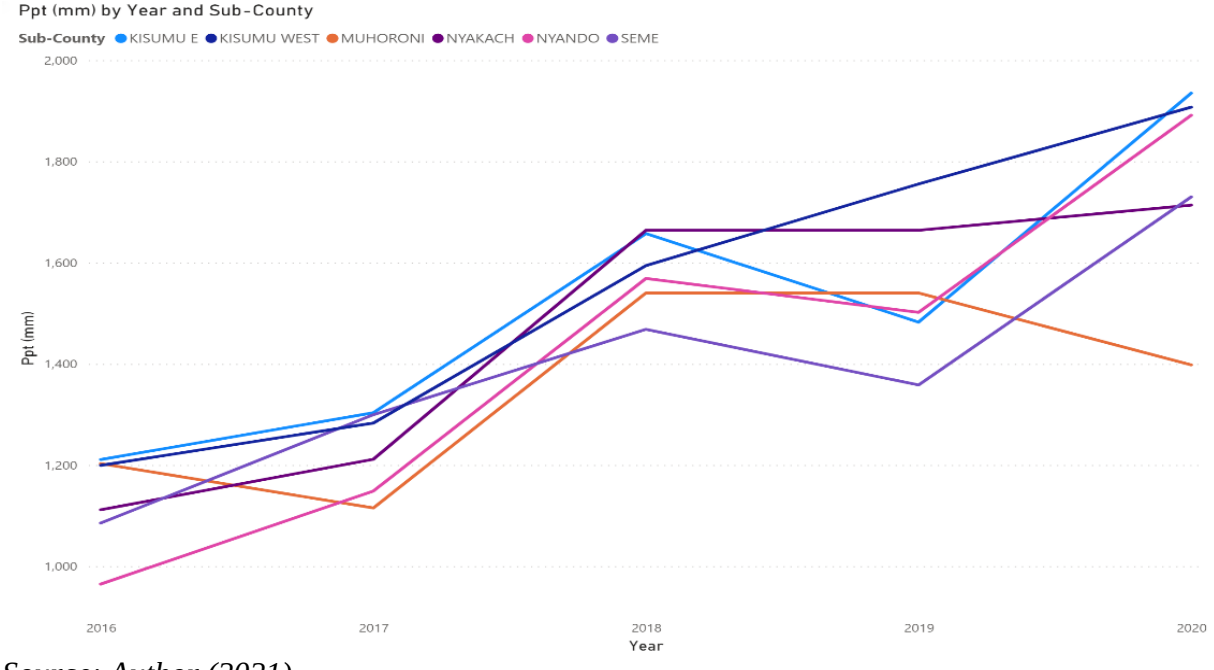


Source: Author (2021)

In looking at the precipitation trends per sub county over the years. A line graph was plotted which shows how different sub counties received the rains as shown in figure 12 below. As the graph showed an increase in rainfall amount for other sub counties, Muhoroni had a drop in rainfall amount in 2020.

FIGURE 12

A Line Graph Showing Precipitation by Year and Sub-County



Source: Author (2021)

4.3 Research Findings

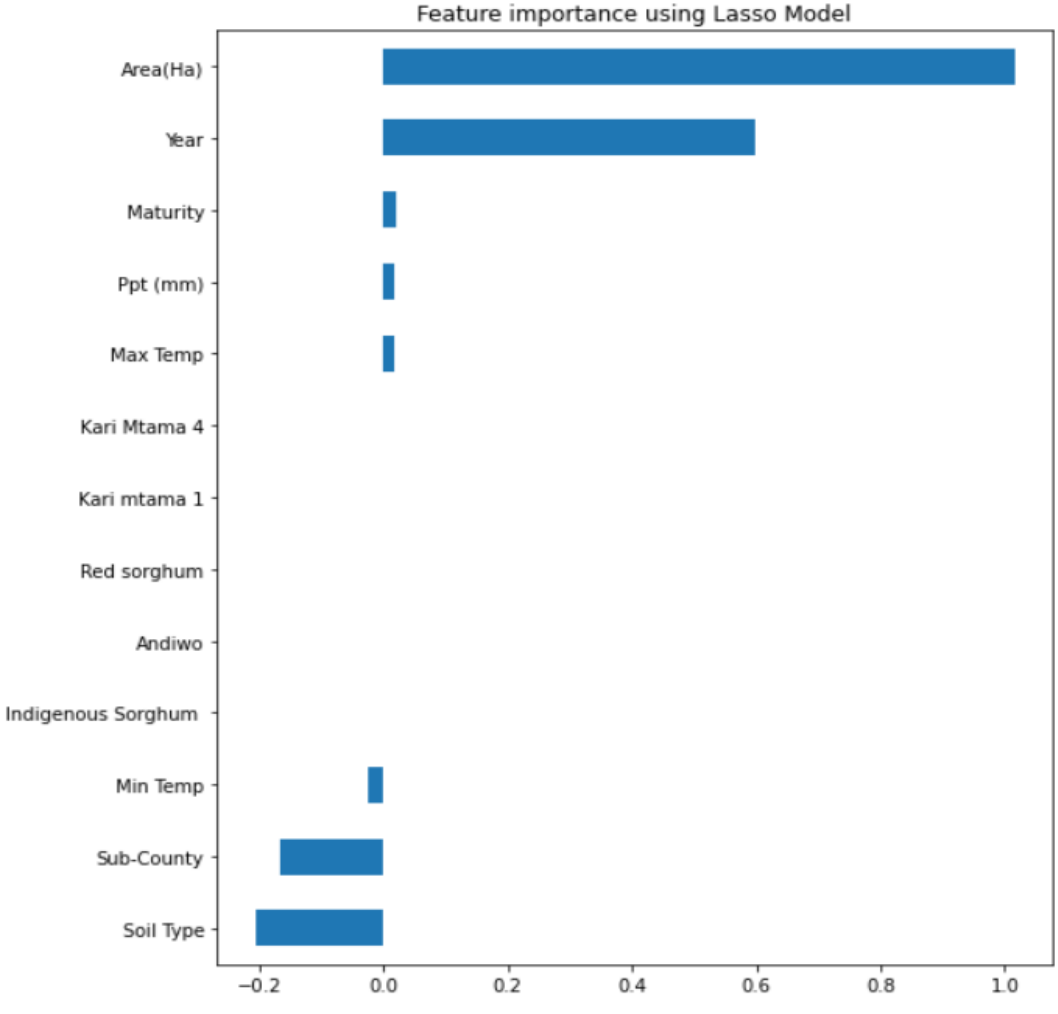
4.3.1 Objective one Results

The first objective of this research was to identify factors affecting sorghum yield in Kisumu County. Feature selection was done on the dataset to identify the relevant factors to be considered for the creation of the model. Feature selection in machine learning is very important because it helps to remove redundant features and reduce the complexity of the model. Only relevant features which have a high impact on the output of the model are picked. There are different feature selection techniques in machine learning broadly classified as filter methods, wrapper methods, embedded methods and hybrid models. This thesis used the embedded method as they offer better accuracy compared to filter and wrapper methods ("Feature selection techniques in machine learning," 2020; Shetye, 2019).

Embedded methods check on the iterations made during the training process of the model and identify features with the most contribution to the training for each iteration. This thesis

adopted a LASSO Regularization which adds a penalty to the different features using a threshold coefficient. Irrelevant features are penalized and their coefficients reduced to zero which are then discarded. Below is figure 13 showing the results of the features that were identified by the Lasso model.

FIGURE 13
Feature Importance Using LASSO Model



Source: Author (2021)

The model identified the following variables as relevant to the yield: Area, Year, Maturity, Precipitation, Maximum temperature, Minimum temperature, Soil type and Sub County.

4.3.2 Objective two Results

The second objective of this research was to develop a deep learning model that predicts the sorghum yield in Kisumu County. In the methodology we stated coming up with two deep

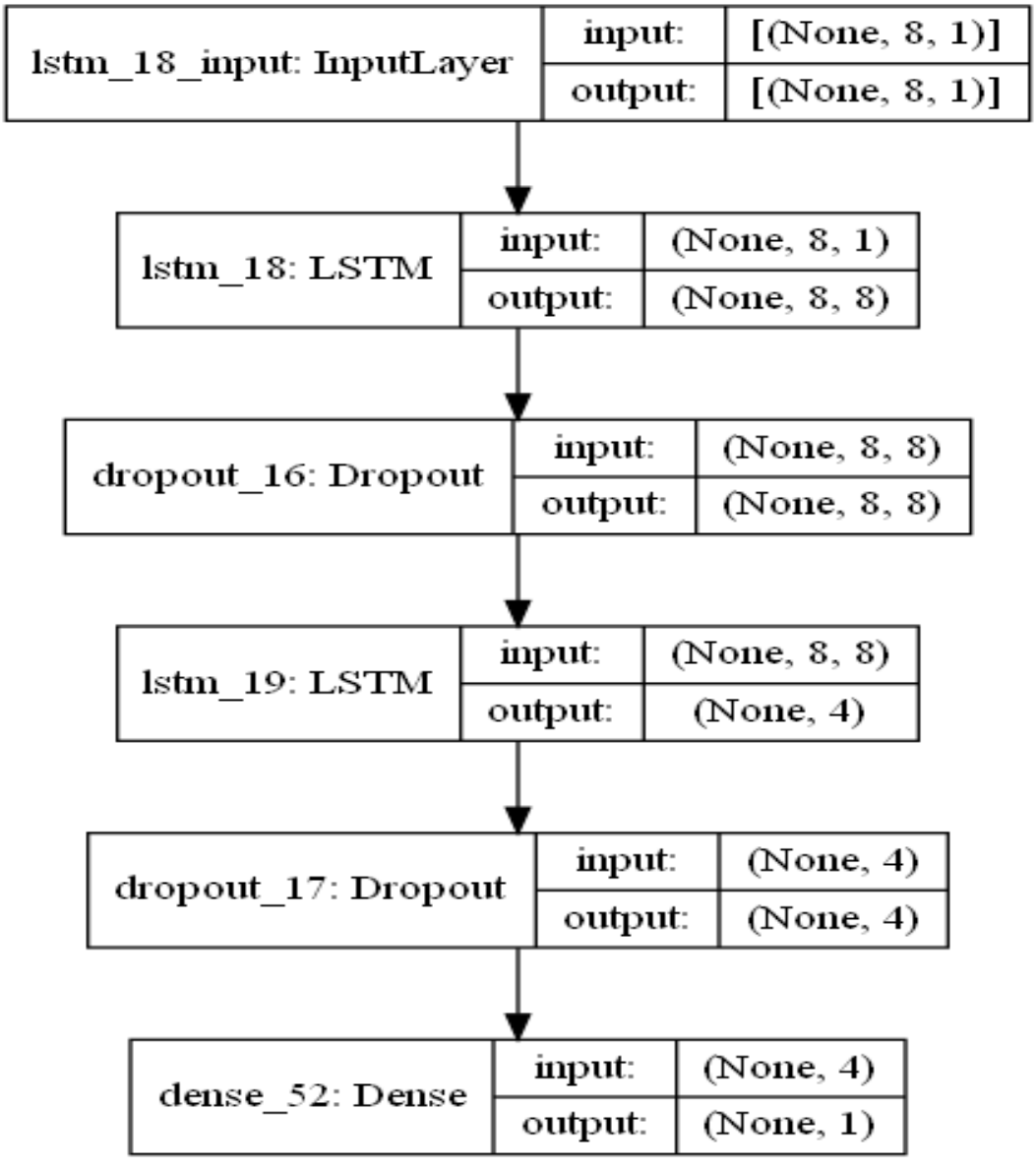
learning models so as to compare their performance and identify the ideal model. The Recurrent Neural Network (RNN) and the Deep Neural Network (DNN) for yield prediction were developed.

4.3.2.1 The RNN Model

The RNN model that was developed was a type called the Long Short-Term Memory (LSTM). This is a type of RNN which has the ability of learning a series of observations. They work well with forecasting time series data. However, overfitting of training data is one of the major issues that affects this model. This issue can be mitigated by introducing a dropout regularization layer. The dropout layer randomly destroys activations by a given factor hence ensuring that the network does not rely on any given activation to be present as they might be squashed at any given moment. So, the model is forced to learn a redundant representation for everything to make sure that at least some of the information remains. This in turn makes the model more robust and prevents overfitting.

The model created is made up of an input layer, two LSTM layers, two dropout layers, and a dense output layer. The parameters specified in the input layer consists of 8 input units, the input shape which was initially in 2-Dimension but reshaped to 3-Dimension since each LSTM layer must be in 3-Dimension. The first LSTM layer consists of 8 units with a consecutive Dropout layer which has a dropout rate of 0.4. The second LSTM layer follows up with 4 units and a consecutive dropout layer with 0.4 dropout rate. The last layer is the dense output layer. The model is compiled with the Adam optimization algorithm and a loss function of Mean Square Error (MSE). Fitting of the training and test data was done with 100 epochs and a batch size of 6. Below is Figure 14 showing a plot of the model described above.

FIGURE 14
Plot Model of RNN



Source: Author (2021)

The total number of trainable and non-trainable parameters of the model can be shown in Figure 15 below which is the model summary. There are a total of 533 trainable parameters, 320 parameters from the first LSTM layer, 208 parameters from the second LSTM layer and 5 from the dense output layer. The number of parameters is obtained from the computation of weights and biases. The dropout layers do not have any parameters since the neurons are randomly dropped-out hence no downstream activation of the neurons during forward pass. This will result in no

updating of weights during the backward pass too. The model summary also shows the layer types and their output shapes.

FIGURE 15
RNN Model Summary

```
regressor.summary()
```

```
Model: "sequential_24"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_18 (LSTM)               (None, 8, 8)              320
_________________________________________________________________
dropout_16 (Dropout)         (None, 8, 8)              0
_________________________________________________________________
lstm_19 (LSTM)               (None, 4)                 208
_________________________________________________________________
dropout_17 (Dropout)         (None, 4)                 0
_________________________________________________________________
dense_52 (Dense)             (None, 1)                 5
=================================================================
Total params: 533
Trainable params: 533
Non-trainable params: 0
_________________________________________________________________
```

Source: Author (2021)

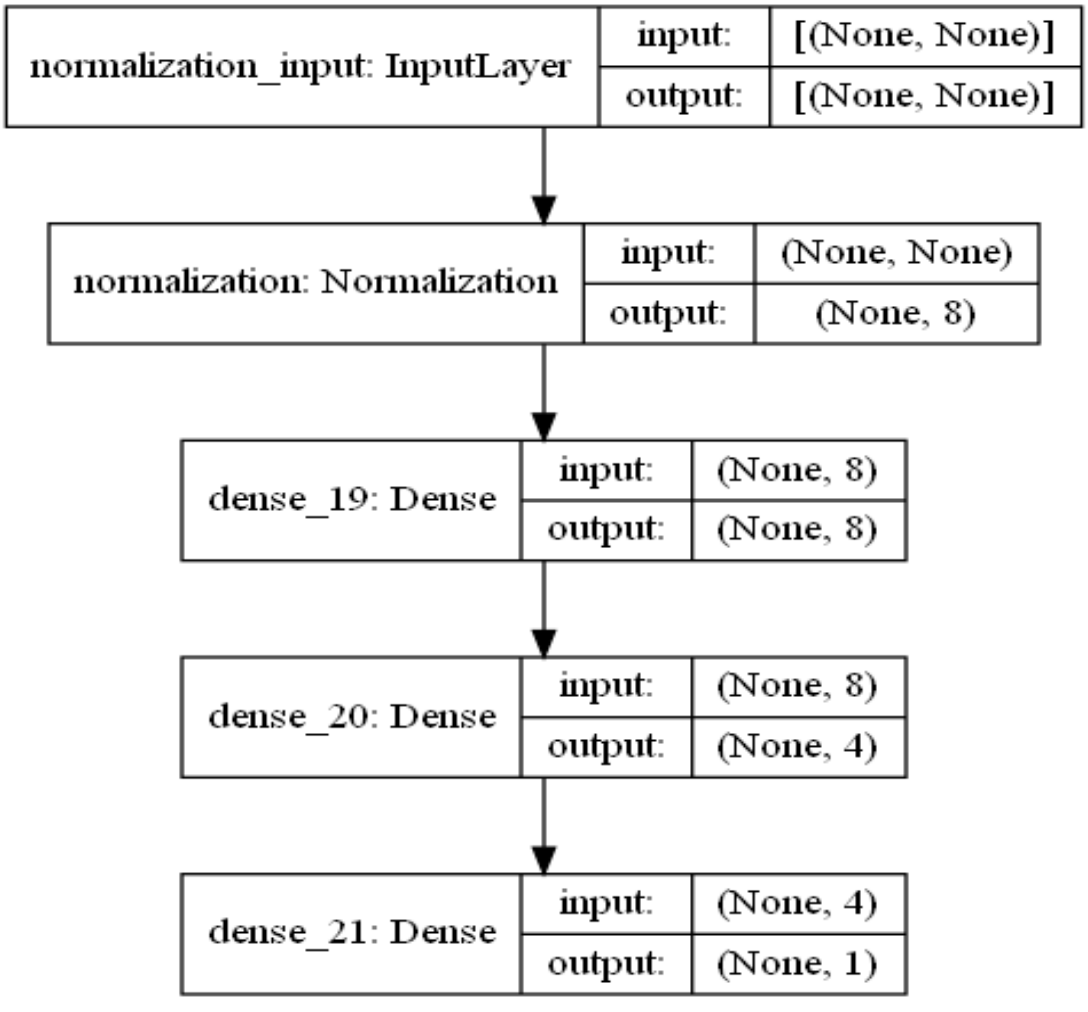
The DNN Model

A DNN model is a type of ANN with multiple layers. An ANN model with more than three layers is considered a DNN. In this thesis we came up with a DNN model with four layers. The first layer is composed of the normalization layer which normalizes the input values and it consists of 8 input units. The second layer is a dense layer which consists of 8 units that is followed by another dense layer with 4 units. The last layer is the dense output layer with 1 unit. The first two dense layers are activated with an activation function known as the Rectified Linear Unit (ReLU). The ReLU helps the network to easily learn the complex patterns within the dataset by solving the vanishing gradient problem. The normalization layer also helps solve the vanishing gradient problem by ensuring that the normalized inputs fall within the desired range to avoid saturation of the activation.

The model was then compiled with the Adam optimization algorithm which had a learning rate of 0.1 and a loss function of Mean Square Error (MSE). Fitting of the training and test data was

done with 100 epochs and a batch size of 6 like the RNN with a validation split of 0.2. Below is Figure 16 showing a plot of the DNN model described above.

FIGURE 16
Plot Model of DNN



Source: Author (2021)

The model summary of the DNN model as shown in Figure 17 below shows the layer types that were involved in the model, the output shapes and the number of parameters that were used. The normalization layer had 17 parameters which were non trainable. The non-trainable parameters represent the number of weights that could not be updated or optimized during training by use of backpropagation. The normalization layer stores the mean and standard deviation for activations that will be used during testing time of the model. The first dense layer contains 72 trainable parameters, the second dense layer contains 26 trainable parameters and the last dense layer with 5 trainable parameters bringing a total of trainable parameters to 113 while the non-trainable parameters at 17. The total number of parameters used in the model is 130.

FIGURE 17
DNN Summary Model

```
dnn_model = build_and_compile_model(normalizer)
dnn_model.summary()
```

```
Model: "sequential_7"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
normalization (Normalization (None, 8)                 17
_________________________________________________________________
dense_19 (Dense)             (None, 8)                 72
_________________________________________________________________
dense_20 (Dense)             (None, 4)                 36
_________________________________________________________________
dense_21 (Dense)             (None, 1)                 5
=================================================================
Total params: 130
Trainable params: 113
Non-trainable params: 17
```

Source: Author (2021)

4.3.3 Objective three Results

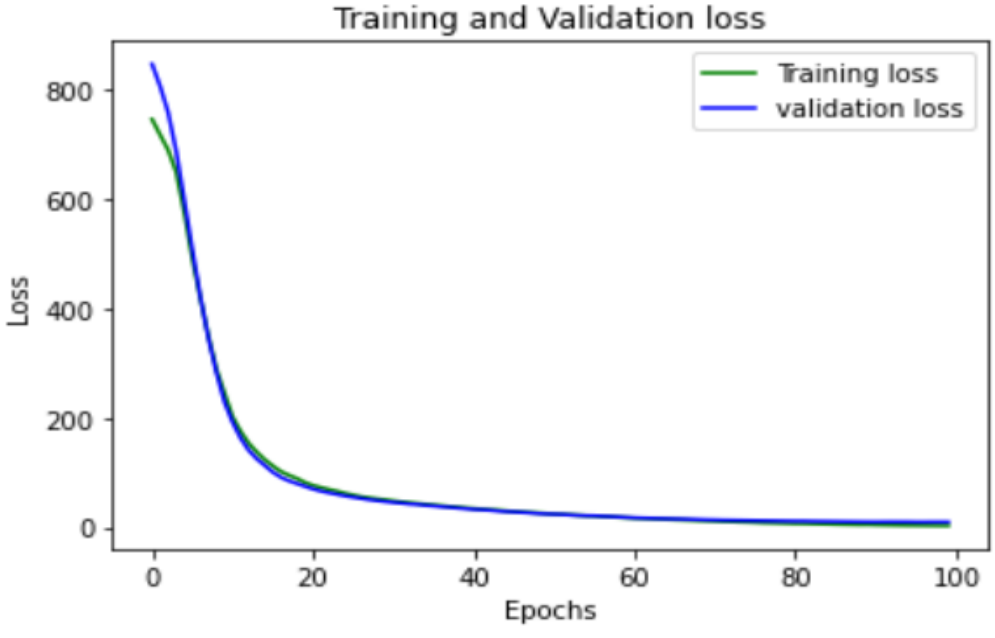
The third objective of this research was to test and validate the prediction model created. A K-fold cross validation technique was used with the Mean Squared Error (MSE) and Squared Mean Squared Error (RMSE) as the metrics for both the RNN and the DNN model. In choosing the value of *k*, 10 is very common and easily recommended in the machine learning field but is being discouraged by James et al. (2013) as the values 10 and 5 tend to excessively suffer from high variance or high bias. An ideal *k* value is that which successfully splits the data into *k* groups with the same number of samples. In this thesis our *k* value was 6.

RNN Model

A learning curve was plotted to identify how a specified metric can learn during the training process of the model. These learning curves can be used to either optimize the model or to check the performance of the model. The learning curve plotted was the train-validation loss over time which consists of a training loss and a validation loss. The training loss shows how well a model fits the training data while the validation loss shows how well the model can fit new data. By monitoring the behavior of the model, changes can be made within the model to ensure that the model is free from high variance which is overfitting of the model and high bias which is the

underfitting of the model. The Figure 18 below shows a plot with variance tradeoff after multiple fine-tuning and optimization of the parameters.

FIGURE 18
Loss Function for RNN Model



Source: Author (2021)

The Root Mean Squared Error (RMSE) and the Mean Squared Error (MSE) which were the chosen evaluation metrics were used and the model achieved a RMSE of 0.54 and an MSE of 0.29 as shown in the Figure 19 below. These results fall within the conventional range of values believed to predict data accurately.

FIGURE 19
RNN model evaluation results

```
accuracy = regressor.evaluate(x_train, y_train)
print('MSE: %.2f' % (accuracy*100))
```

```
2/2 [==============================] - 0s 0s/step - loss: 0.0029
MSE: 0.29
```

```
import math
```

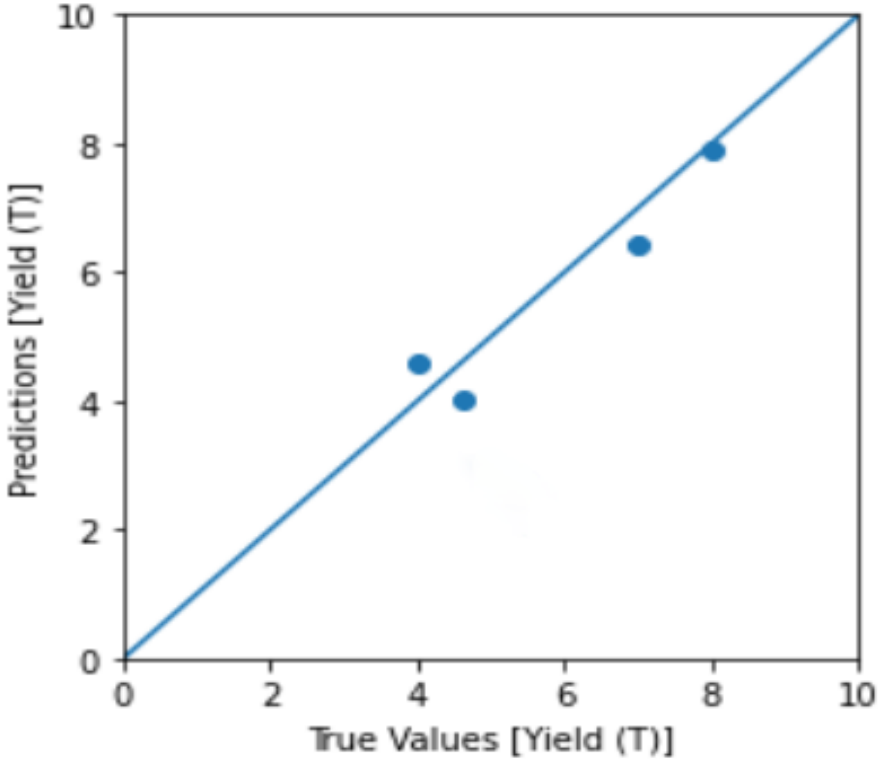
```
rmse = math. sqrt(accuracy*100)
rmse
```

```
0.540770214242929
```

Source: Author (2021)

In an attempt to plot the yield predictions, the following results were obtained as shown in the Figure 20 below. All of the values predicted fell closer to the line of best fit meaning that the model was able to make accurate predictions.

FIGURE 20
Plotting predictions for RNN

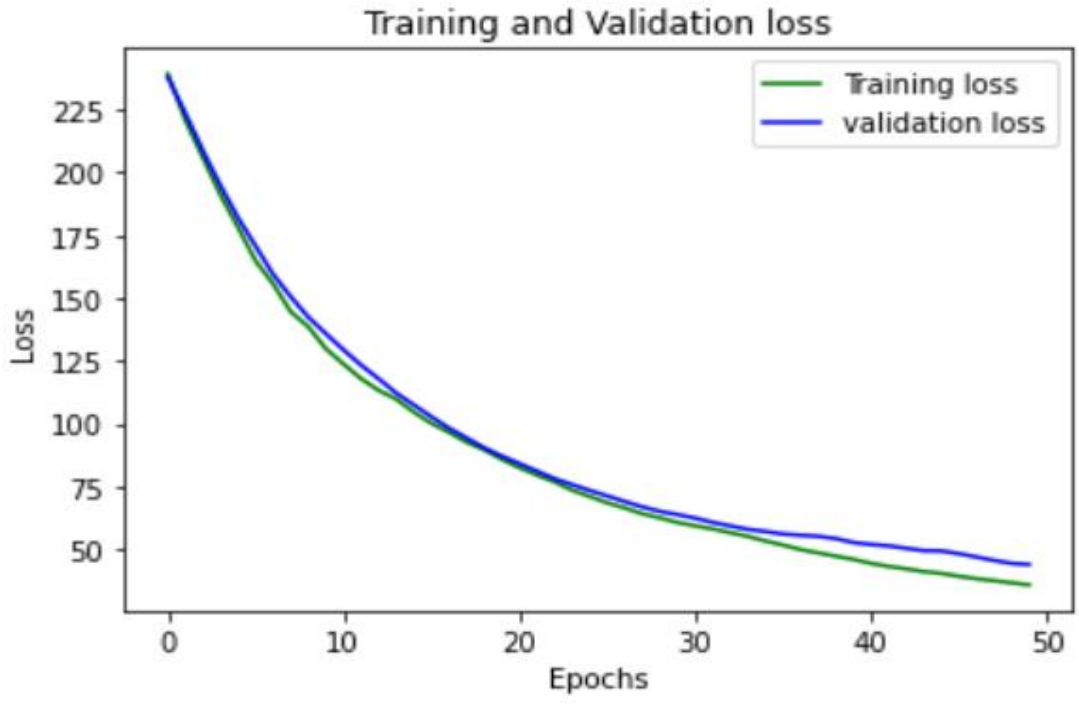


Source: Author (2021)

DNN Model

A train-validation loss over time was also plotted for the DNN model and achieved a variance tradeoff as shown in the Figure 21 below.

FIGURE 21
Loss Function for DNN model



Source: Author (2021)

The DNN model managed a RMSE of 1.17 and a MSE of 1.38 which are relatively minimal errors hence can be used to predict data accurately although they have been less accurate compared to the RNN model. Below is Figure 22 showing the results achieved.

FIGURE 22
DNN Model Evaluation Results

```
print("MSE",np.sqrt(mean_squared_error(y_test,y_pred)))
```

```
MSE 1.3844373104863459
```

```
MSE = np.sqrt(mean_squared_error(y_test,y_pred))
```

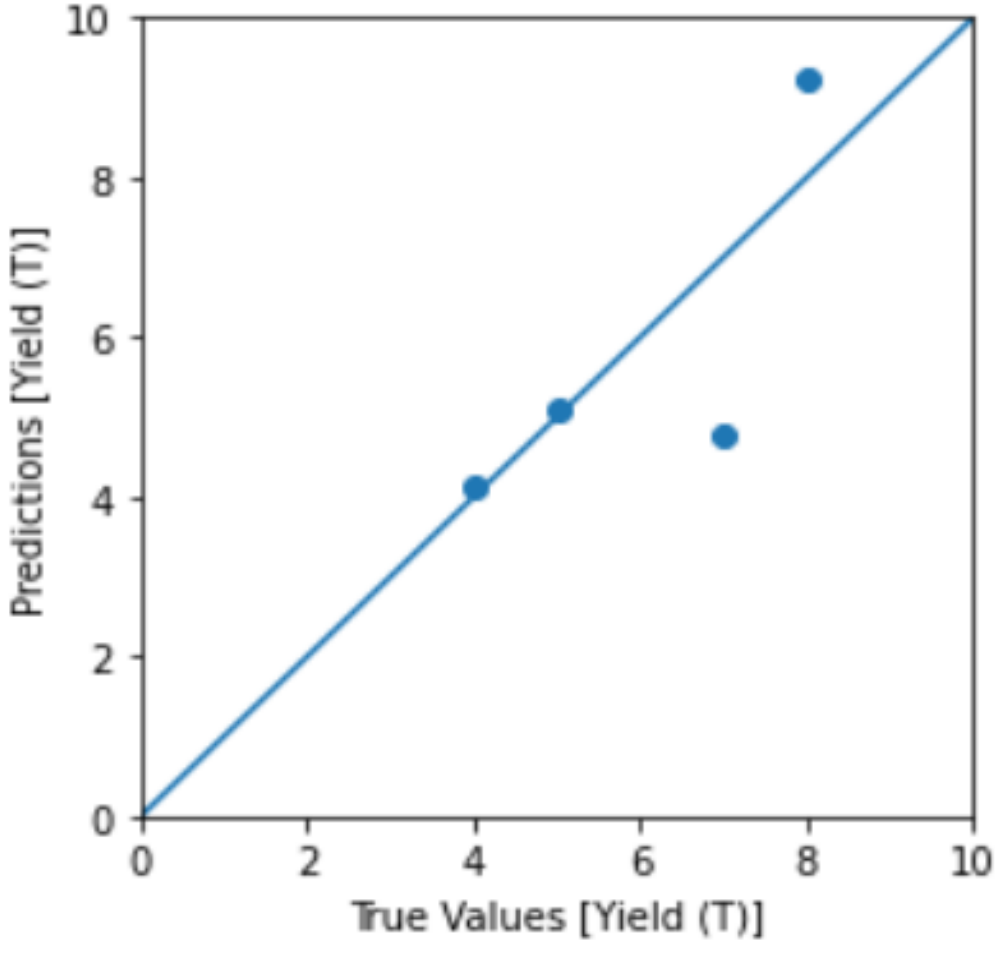
```
rmse = math. sqrt(MSE)
rmse
```

```
1.1766211414411802
```

Source: Author (2021)

In an attempt to also plot the yield predictions of the DNN model, the following results were obtained as shown in the Figure 23 below. Two of the values predicted fell along the line of best fit while the other two were slightly far off on either side.

FIGURE 23
Plotting Predictions for DNN



Source: Author (2021)

4.4 Discussion of Results

The main aim of this study was to come up with a deep learning model for predicting sorghum yield. In order to achieve this, different factors affecting sorghum yield were identified. Through Lasso regularization which is a feature selection technique, 8 factors were found to have significant impact on the sorghum yield. The factors identified were; Sub County, Land area, Year, Time taken for sorghum to mature, Amount of rainfall received in the season, Average maximum and minimum temperatures and soil type. These factors are consistent with findings from the

empirical literature reviews. A systematic literature review by Van Klompenburg et al (2020) on crop yield prediction using machine learning techniques identified the frequency of the most common features used and the most common groups used for crop prediction. The most commonly used features are temperature, soil type, crop information, soil maps and humidity in that order while the most common groups used are soil information, solar information, humidity and nutrients. The following tables 3 and 4 below show the frequencies of the common features used and commonly used grouped features respectfully.

TABLE 3

Common Features Used for Yield Prediction

All features used.

Feature	# of times used
Temperature	24
Soil type	17
Rainfall	17
Crop information	13
Soil maps	12
Humidity	11
pH-value	11
Solar radiation	10
Precipitation	9
Images	8
Area of production	8

Source: (Van Klompenburg et al., 2020)

TABLE 4

Common Grouped Features for Yield Prediction

Grouped features.

Group	# of times used
Soil information	54
Solar information	39
Humidity	38
Nutrients	28
Other	24
Crop information	14
Field management	12

Source: (Van Klompenburg et al., 2020)

Different studies however use different combinations of the above-named features mainly due to the scope of the study and the availability of data. This study managed to pick features from every group that was mentioned in the conceptual framework.

This study came up with two models which were fitted with the same dataset and parameters and evaluated based on their performance. The models that were developed were the RNN – LSTM model and the DNN model. The RNN model outperformed the DNN model by achieving a RMSE value of 0.54 while the DNN model achieved a RMSE value of 1.17 which therefore concluded RNN as the ideal model for sorghum yield prediction. A comparison of results of the two models is consistent with other studies where the two models were used which found out that RNN was performing better than the DNN model. A review on crop predictions using deep learning techniques by Dharani et al. (2021) analyzed the models on the perspective of their abilities to do classification and regression tasks. The DNN model performed with an average accuracy of 60-70% while the RNN achieved an average accuracy of 83-89%. The RNN accuracy can be attributed to its ability to hold data (previous output data for a short period of time) and having a feedback loop.

Van Klompenburg et al (2020) in their study noted that a model having more features did not guarantee the best performance when it came to yield predictions. This study however is inconsistent with the said findings as this model tried to consider all available features and achieved better results compared to the models that were discussed in the empirical literature review. An instance is where Khaki & Wang (2019) used environmental and genotype data to predict corn yield and achieved a RMSE of 11% with DNN model while this study used more features and achieved a RMSE of 1.17% which is far more accurate compared to the other model.

4.5 Summary

In this chapter, secondary data of sorghum yield data was obtained. The data was then taken through a series of data cleaning and preprocessing steps so as to come up with a cleaned dataset. Descriptive analytics was done on the cleaned dataset so as to identify different relationships within the dataset features as well as their relationship towards the target feature which was the sorghum yield. A feature selection was conducted on the dataset to identify the relevant features that had significant impact on the sorghum yield that will be used during the coming up of the deep learning models. Two deep learning models were developed; the RNN model and the DNN model that were used to predict the sorghum yield with the RNN model chosen as the ideal model for sorghum yield prediction. Accurate sorghum yield predictions are crucial

information that will impact various shareholders as previously discussed in the significance of this study.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter deduces conclusions from the previous chapter which was discussing the findings that were derived from the study. It highlights the contribution of the study towards academia and information technology, some of the limitations and challenges that the study experienced and recommendations for future works and research.

5.2 Conclusions

In this study, as it is widely stated, crop yield prediction is not a simple task since multiple factors are put into consideration. Multiple factors were identified and taking a dataset through feature selection is one of the most crucial steps in identifying factors that are of high significance to the identified target feature.

In order to achieve accurate results, data preprocessing is a very important step. This will ensure that the dataset is cleaned by removing null and inconsistent values, the data can be normalized so as to fit to a common scale of use and it can also be encoded into numeric values so as to fit in a machine learning model.

Hyperparameters in any given model determine how well the identified model will perform. These hyperparameters have to be configured before a model is executed. Some of the crucial hyperparameters in a model are; the test split ratio which determines how the data will be split into train and test data, the type of optimization algorithm that will be used like the Adams optimizer and their learning rates, the type of activation function that will be used in a neural network like ReLU, the loss function metric that will be used in a model, the number of activation units a layer will have, the dropout rate , number of epochs required to train a network and the batch size.

Visualization of the created models also helps in monitoring and identifying the behaviors of the models. This helps during the optimization and improvements of the models. Some of the visualizations are like the train-validation loss over time graphs, prediction plots, heatmaps during feature selections and many more.

Evaluation and validation of the models is also helpful in ensuring that the created models achieve optimum results and are able to execute the tasks they were intended to perform.

5.3 Contributions of the study

This study has significantly contributed to the crop yield prediction field by coming up with deep learning prediction models that consider multiple grouped features and provide excellent accuracy rates. It has also helped to identify some of the main features that affect sorghum yield in Kisumu County which can be replicated to other sorghum growing areas.

One of the major issues facing machine learning algorithms is the black box issue whereby the processes that happen between the input up to the output are masked and cannot be explained. A feature selection with lasso regularization comes in handy as it is able to show the strengths of each input.

Through this study, researchers are able to appreciate the fact that python as a programming language is well equipped to perform data analysis from end to end. It is well equipped with libraries that are suitable to handle machine learning tasks like keras and tensorflow. The libraries together with other python libraries can handle tasks from data preprocessing, modeling, evaluation and visualization of the data. This one stop shop eliminates the need to migrate from one platform to the other in different stages of analysis hence improving the overall productivity.

This study can be used as a benchmark for future studies relating to crop yield prediction and specifically to sorghum as a crop.

5.4 Limitations of the study

This study was focused on coming up with a deep learning model for sorghum yield prediction. This involved first identifying factors that affect sorghum yield. Within the group features, some of the features initially identified like size of household and availability of labor in the socio-economic factors could not be factored in as the data represented a whole sub county and not individual farmers hence could not be quantified.

5.5 Recommendations for Future Research

Based on the limitations of the study above, in order to have more features and accuracy in the data, increasing the scope and the size of the research to cover individual farmers rather than grouping farmers into units is recommended.

In order to address the inconsistencies on whether having a model with more features will perform better than a model that has lesser features and vice versa, it is recommended that more research and simulations be done on both models with many features and that with lesser features so as to come up with the right conclusion.

Many crop prediction models have been done using machine learning techniques with little activity done with regards to deep learning models. For future research, more deep learning models should be considered as a topic of research.

Future research should also focus on conducting a series of tests on deep learning algorithms so as to be able to define a range of the size of a dataset in determining their level of performance or identifying what data is good enough to run a particular model.

References

Agarwal, S., & Tarar, S. (2021). A hybrid approach for crop yield prediction using machine learning and deep learning algorithms. *Journal of Physics: Conference Series, 1714,* 012012. https://doi.org/10.1088/1742-6596/1714/1/012012

AI vs. machine learning vs. deep learning vs. neural networks: What's the difference? (2020, May 27). IBM - United States. https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks

Artificial neural network (ANN). (n.d.). Investopedia. https://www.investopedia.com/terms/a/artificial-neural-networks-ann.asp

Bandaru, V., Stewart, B. A., Baumhardt, R. L., Ambati, S., Robinson, C. A., & Schlegel, A. (2006). Growing Dryland grain sorghum in clumps to reduce vegetative growth and increase yield. *Agronomy Journal*, *98*(4), 1109-1120. https://doi.org/10.2134/agronj2005.0166

Brownlee, J. (2016). Deep learning with Python: develop deep learning models on Theano and TensorFlow using Keras. Machine Learning Mastery. https://www.coursehero.com/file/32130187/deep-learning-with-pythonpdf/

Brownlee, J. (2019). *Deep learning for computer vision: Image classification, object detection, and face recognition in Python*. Machine Learning Mastery.

Chimoita, E. L., Onyango, C. M., Gweyi-Onyango, J. P., & Kimenju, J. W. (2017). Factors Influencing Uptake of Improved Sorghum (Sorghum Bicolor) Technologies in Embu County, Kenya. http://erepository.uonbi.ac.ke/handle/11295/101853

Curran, S., & Cook, J. (2009). Gender & Cropping in Sub-Saharan Africa: Sorghum. *Evans School Policy Analysis & Research Group (EPAR)*. https://epar.evans.uw.edu/research/gender-cropping-sub-saharan-africa-sorghum

Data science methodology and approach. (2019, October 3). GeeksforGeeks. https://www.geeksforgeeks.org/data-science-methodology-and-approach/

Demissie, A. (2013). Determinants of income diversification among rural households: The case of smallholder farmers in Fedis district, eastern hararghe zone, Ethiopia. *Journal of Development and Agricultural Economics*, *5*(3), 120-128. https://doi.org/10.5897/jdae12.104

Dharani, M. K., Thamilselvan, R., Natesan, P., Kalaivaani, P., & Santhoshkumar, S. (2021). Review on crop prediction using deep learning techniques. *Journal of Physics: Conference Series*, *1767*(1), 012026. https://doi.org/10.1088/1742-6596/1767/1/012026

Duku, C., & Groot, A. (2020). *Climate change risks and opportunities in sorghum production in Kenya*. https://www.researchgate.net/publication/343569342_Climate_change_risks_and_opportunities_in_sorghum_production_in_Kenya

East African Breweries Limited. (2019). *EABL Annual Report - 2019*. https://www.eabl.com/sites/default/files/eabl-annual-report-2019.pdf

Elavarasan, D., & Vincent, P. M. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. *IEEE Access*, *8*, 86886-86901. https://doi.org/10.1109/access.2020.2992480

Encyclopedia of environmental health. (n.d.). ScienceDirect.com | Science, health and medical journals, full text articles and books. https://www.sciencedirect.com/referencework/9780444639523/encyclopedia-of-environmental-health

Feature selection techniques in machine learning. (2020, December 2). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/

Fu, Y. (2011). Data mining models and tasks. https://academic.csuohio.edu/fuy/Pub/pot97.pdf

Gebbers, R., & Adamchuk, V. I. (2010). Precision agriculture and food security. *Science*, *327*(5967), 828-831. https://doi.org/10.1126/science.1183899

GoK. (2015). *Economic Review of Agriculture (ERA). Ministry of Agriculture Livestock and Fisheries (MoALF)*. Government of Kenya (GoK).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Grand View Research. (2021). *Precision Farming Market Size, Share & Trends Analysis Report By Offering, By Application (Yield Monitoring, Weather Tracking, Field Mapping, Crop Scouting), By Region, And Segment Forecasts, 2021 - 2028* (GVR-1-68038-376-8).

Grossfeld, B. (2020, January 23). *Deep learning vs. machine learning: What's the difference?* Zendesk. https://www.zendesk.com/blog/machine-learning-and-deep-learning/

Guide to farm sorghum in Kenya. (n.d.). Value Farming. https://value.co.ke/article/guide-farm-sorghum-kenya

Holzman, M. E., Carmona, F., Rivas, R., & Niclòs, R. (2018). Early assessment of crop yield from remotely sensed water stress and solar radiation data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *145*, 297-308. https://doi.org/10.1016/j.isprsjprs.2018.03.014

Hong, S., & Hanson, H. (2016). Scaling up Agricultural Credit in Africa. https://assets.ctfassets.net/5faekfvmlu40/7jVMoWYlSoW84GAMayqc4o/0d807

287c05eb6534ad47680089e3458/1af_scaling_up_agricultural_credit_in_africa_2-25_final_v1.pdf

Huntington, T., Cui, X., Mishra, U., & Scown, C. D. (2020). Machine learning to predict biomass sorghum yields under future climate scenarios. *Biofuels, Bioproducts and Biorefining*, *14*(3), 566-577. https://doi.org/10.1002/bbb.2087

Irrigation in sorghum. (n.d.). agropedia |. https://agropedia.iitk.ac.in/content/irrigation-sorghum

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer Science & Business Media.

KALRO. (n.d.). *Sorghum | Agricultural mechanization research institute*. Kenya Agricultural & Livestock Research Organization. https://www.kalro.org/amri/?q=node/166

Karagiannakos, S. (2020, February 26). *Deep learning algorithms - The complete guide*. AI Summer. https://theaisummer.com/Deep-Learning-Algorithms/

Kenya Climate Smart Agriculture Project. (2018). *KCSAP PROJECT IMPLEMENTATION MANUAL*. https://www.kcsap.go.ke/wp-content/uploads/2019/02/Project-Implementation-Manual-PIM.pdf

Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.00621

Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A CNN-RNN framework for crop yield prediction. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.01750

Mmbando, F. E., & Baiyegunhi, L. J. (2016). Socio-economic and institutional factors influencing adoption of improved maize varieties in Hai district, Tanzania. *Journal of Human Ecology*, *53*(1), 49-56. https://doi.org/10.1080/09709274.2016.11906955

Muui, C., Muasya, R. M., & KIrubi, D. T. (2013). Baseline survey on factors affecting sorghum production and use in eastern Kenya. *African Journal of Food, Agriculture, Nutrition and Development*, *13*(01), 7339-7342. https://doi.org/10.18697/ajfand.56.11545

Nairobi123. (2013, July 22). *File:Kisumu County location map*. https://commons.wikimedia.org/wiki/File:Kisumu_County_location_map.png. File:Kisumu County location map

Nasir, Vahid & Sassani, Farrokh. (2021). A review on deep learning in machining and tool monitoring: methods, opportunities, and challenges. The International Journal of Advanced Manufacturing Technology. 115. 10.1007/s00170-021-07325-7.

NetSuite.com. (2020, September 23). *Predicting your business's future*. Oracle NetSuite. https://www.netsuite.com/portal/resource/articles/financial-management/predictive-modeling.shtml

Njagi, T., Onyango, K., Kirimi, L., & Makau, J. (2019). *Sorghum Production in Kenya: Farm-level Characteristics, Constraints and Opportunities*. https://www.tegemeo.org/images/_tegemeo_institute/downloads/publications/technical_reports/tr34%20sorghum%20production%20in%20kenya%20farm-level%20characteristics,%20constraints%20and%20opportunities.pdf

Ogeto, R., Cheyuiyot, E., Mshenga, P., & Onyari. (2013). Sorghum production for food security: A socioeconomic analysis of sorghum production in Nakuru County, Kenya. *African Journal of Agricultural Research*. https://academicjournals.org/article/article1386078688_Ogeto%2520et%2520al.pdf

Ogutu, G. E., Franssen, W. H., Supit, I., Omondi, P., & Hutjes, R. W. (2018). Probabilistic maize yield prediction over East Africa using dynamic ensemble seasonal climate

forecasts. *Agricultural and Forest Meteorology*, *250-251*, 243-261. https://doi.org/10.1016/j.agrformet.2017.12.256

Okeyo, S. O., Ndirangu, S. N., Isaboke, H. N., Njeru, L. K., & Omenda, J. A. (2020). Analysis of the determinants of farmer participation in sorghum farming among small-scale farmers in Siaya County, Kenya. *Scientific African*, *10*, e00559. https://doi.org/10.1016/j.sciaf.2020.e00559

Onono, P. A. (2018). Response of sorghum production in Kenya to prices and public investments. *Sustainable Agriculture Research*, *7*(2), 19. https://doi.org/10.5539/sar.v7n2p19

Patel, A. (2019, August 9). *Data science methodology 101*. Medium. https://medium.com/ml-research-lab/data-science-methodology-101-2fa9b7cf2ffe

Pai, A. (2020, October 19). *CNN vs. RNN vs. ANN - Analyzing 3 types of neural networks in deep learning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/

Sharma, G. (2021, May 27). *Regression algorithms | 5 regression algorithms you should know*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/5-regression-algorithms-you-should-know-introductory-guide/

Shetye, A. (2019, February 12). *Feature selection with sklearn and pandas*. Medium. https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b

Shook, J., Wu, L., Gangopadhyay, T., Ganapathysubramanian, B., Sarkar, S., & Singh, A. K. (2018). Integrating genotype and weather variables for soybean yield prediction using deep learning. https://doi.org/10.1101/331561

Singh, A., Ganapathysubramanian, B., Singh, A. K., & Sarkar, S. (2016). Machine learning for high-throughput stress Phenotyping in plants. *Trends in Plant Science*, *21*(2), 110-124. https://doi.org/10.1016/j.tplants.2015.10.015

Socio-economic and ecological characteristics. (n.d.). Home | Food and Agriculture Organization of the United Nations. https://www.fao.org/3/AB396E/ab396e02.htm

Sorghum production and area by counties 2018.csv. (n.d.). Welcome - Kilimo Open Data. https://kilimodata.developlocal.org/dataset/kenya-sorghum-production-by-counties/resource/057639a0-53e8-4922-a0b5-7b5aab08d2cb

Sorghum. (n.d.). Infonet Biovision Home. https://infonet-biovision.org/PlantHealth/Crops/Sorghum

Sridhara, S., Ramesh, N., Gopakkali, P., Das, B., Venkatappa, S., Sanjivaiah, S., Kumar Singh, K., Singh, P., El-Ansary, D., Mahmoud, E., & Elansary, H. (2020). Weather-based neural network, stepwise linear and sparse regression approach for rabi sorghum yield forecasting of Karnataka, India. *Agronomy*, *10*(11), 1645. https://doi.org/10.3390/agronomy10111645

Supervised vs. unsupervised learning: What's the difference? (2021, March 12). IBM - United States. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

Suvedi, M., Ghimire, R., & Kaplowitz, M. (2017). Farmers' participation in extension programs and technology adoption in rural Nepal: A logistic regression analysis. *The Journal of Agricultural Education and Extension*, *23*(4), 351-371. https://doi.org/10.1080/1389224x.2017.1323653

USAID. (2013). *KENYA AGRICULTURAL VALUE CHAIN ENTERPRISES PROJECT (USAID-KAVES)*. https://pdf.usaid.gov/pdf_docs/pa00m2s4.pdf

Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709. https://doi.org/10.1016/j.compag.2020.105709

Wang, C. (2019, January 8). *The vanishing gradient problem*. Medium. https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484

Wanyama, R., Mathenge, M. W., & Mbaka, Z. S. (2016). Agricultural Information Sources and their Effect on Farm Productivity in Kenya. http://www.renapri.org/wpcontent/uploads/2017/01/Tegemeo_WP62_2016.pdf

World Bank Group. (2015). *Ending poverty and hunger by 2030: An agenda for the global food system*. https://documents1.worldbank.org/curated/en/700061468334490682/pdf/95768-REVISED-WP-PUBLIC-Box391467B-Ending-Poverty-and-Hunger-by-2030-FINAL.pdf

World Bank Group. (2016). *Poverty and shared prosperity 2016: Taking on inequality*. Poverty and Shared Prosperity. https://openknowledge.worldbank.org/handle/10986/25078

World Bank Group. (2017). *Kenya economic update, December 2017: Poised to bounce back?* https://openknowledge.worldbank.org/handle/10986/29033

World Bank Group. (2019). *Kenya economic update, April 2019, No. 19: Unbundling the slack in private sector investment - Transforming agriculture sector productivity and linkages to poverty reduction*. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/820861554470832579/kenya-economic-update-unbundling-the-slack-in-private-sector-investment-transforming-agriculture-sector-productivity-and-linkages-to-poverty-reduction

World Bank Group. (n.d.). *Capacity Needs Assessment for Improving Agricultural Statistics in Kenya*. https://documents1.worldbank.org/curated/zh/801111542740476532/Capacity-Needs-Assessment-for-Improving-Agricultural-Statistics-in-Kenya.docx

Yadav, R., Yadav, S., Gunjal, N., & Mandal, S. (2019). Agricultural Crop Yield Prediction using Deep Learning Approach. *International Research Journal of Engineering and Technology (IRJET)*, *6*(12).

Zannou, J. G., & Houndji, V. R. (2019). Sorghum yield prediction using machine learning. *2019 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART)*. https://doi.org/10.1109/biosmart.2019.8734219

APPENDICES

Appendix I: Research Schedule

TABLE 5

Research Schedule

Activities	**Task**	**2021**											
		1	2	3	4	5	6	7	8	9	10	11	12
Proposal development	Literature review												
	Proposal writing												
	Proposal presentation, correction and defense												
Project identification, development and identification	Data collection												
	Analysis of collected data												
	Project Writing												
	Project presentation, corrections and defense												

Appendix II: Resources and Budget

TABLE 6

Resources and Budget

Budget item	**Total (Ksh)**
Laptop – High performance computer (AMD Ryzen 7)	150,000
Internet cost	20,000
Miscellaneous	50,000
Ksh.220,000	