International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

# Role of Big Data in Agriculture

**Madhuri J, Indiramma M**

***Abstract- Big Data analytics in the agricultural sector has huge potential to cater to requirements of food production. This review highlights the role of Big Data in pertinent data acquisition from factors affecting the agriculture such as weather, soil, diseases, remote sensing and the prospects of agricultural data analysis towards smart farming. Incorporating modern technologies in farming practices continuously monitor the environment, thus producing large quantity of data. Hence there arises the need for advanced practical and systematic strategies to correlate the different factors driving the agriculture to derive valuable information out of it. Big Data can be a promising aspect for the future of food production and sustainability of agriculture. Leveraging big data in the agricultural sector can provide insights in farming practices, helps in making real-time decisions and motivates in incorporating new methods of farming operations. The main objective of this paper is to provide insights into different factors that contribute to making timely recommendations to farmers with regard to smart agricultural techniques.***

***Keywords. Big Data; Agriculture; Smart farming; Soil; Climate; Remote Sensing; IOT.***

## I. INTRODUCTION

Agriculture is considered as the basic human activity for survival [1].Crop yield and food production are dependent on multiple factors. Increase in the global population has raised the food demand and the concerns about food security needs to be quickly addressed [5]. According to 2011 census in India, an estimate of 67.5% of the Indian population is dependent on agriculture. Agricultural sector places a pivotal role in the Indian economy and contributes 17.9% to total Gross domestic product (GDP) composition in India according to the Ministry of Statistics and programmed Implementation, Gov. of India. It has been projected that India's population would reach 1.5 billion by 2035. The surge in population has resulted in narrowing down of man to land ratio critically. It has been witnessed that per capita land availability in India has decreased from .48ha in 1951 to 0.26 ha in 1981 and has further declined to 0.14 ha during 2000. It may taper down to 0.09ha by 2020. Hence the need for accurate prediction of crop yield and food production to small farms at individual farm level is the matter of utmost importance.

**Revised Manuscript Received on December 30, 2019.**
***** Correspondence Author
**Madhuri J***, Research Scholar, Department of Computer Science and Engineering, B.M.S. College of Engineering, Bangalore, India.
**Indiramma M**, Professor, Department of Computer Science and Engineering, B.M.S. College of Engineering, Bangalore, India.



The cropping patterns and the associations with reference to individual land holdings have to be analyzed for suitable crop recommendations thereby increasing crop sustainability and productivity.[40] It is also necessary to perceive how crop growth is sensitive to climate factors, soil conditions, and farming practices. But the farmers in the developing countries face several challenges due to inadequate knowledge in nutrient management in soil, outdated farming activities that have led to overproduction or underproduction of the food products. Farmers are unable to sell their harvest when they don't get the yield as planned qualitatively and quantitatively. This demands focused attention on adapting modern tools in farming and intuitive planning of farm management strategies with analysis of historical data to substantiate the decisions.The digital revolution in agriculture is focused on low-cost data collection of soil conditions, weather station and data collected by the satellites [53]. The farming methods are now supported by innovative digital technologies such as rain gauge, soil sensors, moisture sensors, thermostats with assistance from the Internet of things (IoT). IOT enables the devices to communicate with each other and interact to automate farming practices. Using technology over agricultural land is termed as smart farming. Inclusions of smart machines and sensors in the farms contribute to data growth in the sector. Traditional agriculture systems assume that the parameters of the field are consistent, hence the fertilizer management, irrigation methods, pesticide application, cropping patterns which are not suitable for existing conditions are followed. Using precision agriculture is the concept that enables farmers to understand their crops at the micro level and manage the crops smartly.[14]

### 1.1 Precision agriculture

Precision Agriculture (PA) or precision farming is the strategic trend in developed countries. Precision farming is the site-specific crop managing strategy based on spatial and time-based characteristics of the crop, weather and soil factors within the field. The objective of precision farming is to tailor cropping patterns, fertilizing, irrigation and insecticides in every farm. Farmers who have adapted data-driven farming processes depending on the soil characteristics and weather parameters have accounted for a remarkable rise in the final yield [11][6]. Applying PA strategies in developing countries is a challenge with a limited amount of land and resources. These technologies are used to make decisions on crop management. PA technologies can be categorized into soft PA and hard PA. Soft PA is subject to making farming decisions based on observing soil nature and crops and making decisions based on human experience.

Hard PA is building decision-based on scientific and statistical analysis with full fledge inclusion of technologies such as sensors, GPS, etc. [35] Farmers in developing countries evade the use of PA as they lack skills to manage the PA tools and are financially deficient to make the purchases. Trials, demonstrations, building the confidence about ease of use and its effectiveness would enhance the rate of adaptability to PA [54].

Implementing PA generates huge amounts of diverse data from the agricultural sector. Data acquisition include soil characters, seeding rates, crop yields that can be combined with the historical records such as weather patterns, topography and crop performance [6]. Big data is a potent platform to store the diverse data collected and analyze the data to make site-specific decisions

## II. METHODOLOGY

This review considers the areas majorly influencing the crop yield and use of big data technologies in agriculture. Climate, diseases, and pests, soil mapping, use of IoT are the areas considered. Firstly the conference and journal work related to each factor is searched from the research papers databases such as IEEE, ScienceDirect, Researchgate were searched using the web indexing service Google scholar. Secondly, most relevant literature were filtered and their effects on the crop recommendations were summarized. This study analyses the agriculture factor considered, the problem addressed and solution provided to overcome the problem in each area. Thirdly the literature related to the use of big data in agriculture were searched to study the actual application of big data analytics in agricultural practices. Section 3 states the reasons for applying big data in the field of agronomy and the potential benefits of big data. Section 4 summarizes the major data sources that contribute to research related to crop recommendation, management, and yield prediction. The most relevant papers are summarized in Table1, Table2, and Table3. Section 5 summarizes the role of machine learning in big data analytics. Section 6 addresses the challenges faced to apply big data in agriculture.

## III. BENEFITS OF USING BIG DATA IN AGRICULTURE

The term agricultural "big data" helps to realize the necessity of considerable investments in infrastructures for storage and processing of agricultural data. Big data promises precision data storage, processing and analyzing that was not possible before with traditional methods. It enables searching, aggregating, relating different agricultural datasets to get optimum conclusions in farming. Relating factors such as remotely sensed data (crop health, Leaf Area Index, soil mapping, etc.) with the statistical data (rainfall, temperature, and previous yields) supports the decisions such as crop recommendations, yield prediction, fertilizer recommendation, pest management, forecasting prices, and policy recommendation. [37]

### A Big Data

Big data is the combination of structured and unstructured data that is growing at an unpredictable rate. Big data refers to the enormous amount of digital data that is difficult to manage and analyze using customary software tools and technologies [20]. Big Data is often used to describe a modern trend in which the combination of technology and advanced analytics creates a new way of processing in a useful way. Substantial amounts of data generated by social media, cellphones, and other digital communication tools, contribute to big data. This data is raising enormously and has become tough for capturing, arranging, storing, managing, sharing, analyzing, and visualizing via classic or traditional database software tools. Big data demands a revolutionary move away from classical data analysis. For example, Gartner, Inc. defines big data in similar terms: "Big data is considered as high voluminous, high velocity and high variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."[12]

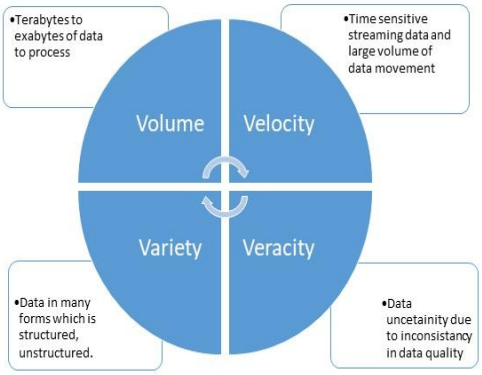


**Figure 1: 4 Vs of Big Data**

Big Data can be characterized by 4 Vs [12]. Volume refers to huge amount of data being generated through a wide range of sources. Velocity refers to the data which are time-sensitive and needs to be collected, stored, processed, analyzed, and acted on quickly. Variety of data refers to the data arising from multiple formats such as structured data in the traditional database, and unstructured text documents; email, video, audio, and financial transactions. Veracity is another term that characterizes big data signifying the data flows that vary greatly with periodic peaks and baseline. This variation may be due to daily, seasonal and event-triggered peak data loads, social media trends and other factors. [20] With this definition, characteristics of big data may be outlined as four Vs., i.e., Volume (large data), Variety (various modalities), Velocity (rapid generation), and Veracity (data inconsistency) The key challenges in the development of big data applications can be listed as [21] Data demonstration: The agricultural datasets have remarkable levels of heterogeneity in terms of data types, organization, semantics, structure, granularity, and ease of access. Data demonstration is the phase which makes the data more significant for the computer to analyze and interpret the results.

*Duplicate reduction*: The datasets can be highly redundant due to duplicate entries. Reducing the duplicate entries contributes to reduction in the size of data but the process has to be carried out without affecting the potential values of the entire dataset.

*Analysis methods*: The Big data analysis system shall process masses of data with limited time or late results shall not be valid in real-time applications.
*Data confidentiality*: The big data service providers or owners are presently depend on external tools to store and analyze the data as maintaining large data sets with limited capacities. This reliability increases the potential safety risks.
*Data scalability*: Big data analytical system must support the expanding datasets in the future. Big data applications must be able to store and process more expanded complex datasets.
Big data analytics is effectively being leveraged in different sectors, such as banking sector, retail sector, insurance, online user behavior recognition, and medicine field. The governmental bodies have initiated the use of big data analysis to address the challenges faced by the country citizens, job opportunities, medical coverage, natural disaster management, and infiltration.

### B. Agricultural Data

Agricultural data collected can be categorized as machine-generated data, the process generated data and human sourced. [2]
*Machine generated* data includes data from sensors, unmanned aerial vehicles, GPS. These data from new technologies may vary from sounds to images.
*Process generated* data includes data collected from farms such as information on planting, monitoring, and recording of the farming process such as seeding fertilizer application.
*Human-sourced* is previously recorded human experiences sourced. Experiences which were previously stored as books are digitized and stored to provide accessibility.
Collecting agriculture data precisely is critical as the data varies from ecology to human, geography to the economy. Agricultural yield is closely related to geographic location termed as spatial data and it is stored as coordinates to locate an area. Crops various stages such as seeding, fertilizing, pest management, weeding, water supply, harvesting. Acquiring data in time is significant for analysis and management. [9]

### C. Big data analysis

Big data analysis can be applied in agriculture with the support of experienced agricultural experts. Experts may include farmers, agro researchers, agro market analysts, distribution specialists, and so on. Experts can provide better and intelligent decision and monitoring the raw material which is nothing but data/Information [37]. The large voluminous agricultural data available creates a necessity to use Big Data analytics to extract valued information and generate accurate results which help in precise decision making. The big data analytics platforms include cloud computing, Internet of Things, machine learning and artificial intelligence. [43] "Big data analysis" is the term used to refer the innovative practices to handle large data [60] [44]. The farmers and related organizations can extract economic value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis of the farming information [61] [62].
Big Data in agriculture can be generally divided into four stages: (i) Data capture (ii) Data storage (iii) Data transformation (iv) Data analysis (v) Data marketing.

Data capture is extracting massive amounts of heterogeneous data from diverse sources like weather stations, remote mapping, cropping patterns, field characteristics, etc. available from the department of agriculture, Indian government.
Data storage is the key challenge in massive data-driven applications, currently, NoSQL technologies are gaining popularity.
Data transformation or the preprocessing is the stage where multiple data representing the same parameter has to be integrated into a single consistent representation.
Data analytics derives the value out of the captured data and interprets the results converting it into information.
Data marketing is connecting the analytical results achieved to make decisions about the processes to be followed.

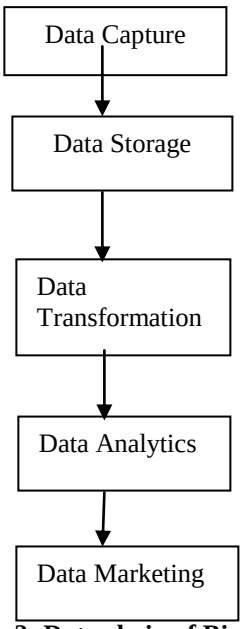


**Figure 2: Data chain of Big Data applications**

## IV. BIG DATA FRAMEWORK WITH FACTORS INFLUENCING AGRICULTURE

Different agricultural issues suitable crop selection, irrigation methods, fertilizer selection, and yield prediction can be addressed by leveraging big data technologies.
Based on the data chain of big data applications shown in Figure 2, the following framework proposes a system that can generates mobile In agriculture, a variety of data as shown in Figure 3 originates from the following sources
Soil: Agricultural agencies such as NBSS and LUP (National Bureau of soil survey and land use planning), satellite images.
Climate: Meteorological departments, government organizations, Historical data
Pests: Images, Historical data.
Remote Sensing: Satellite images from Government organizations such as ISRO
IOT: Sensors, Radiofrequency identifiers(RFID) etc.

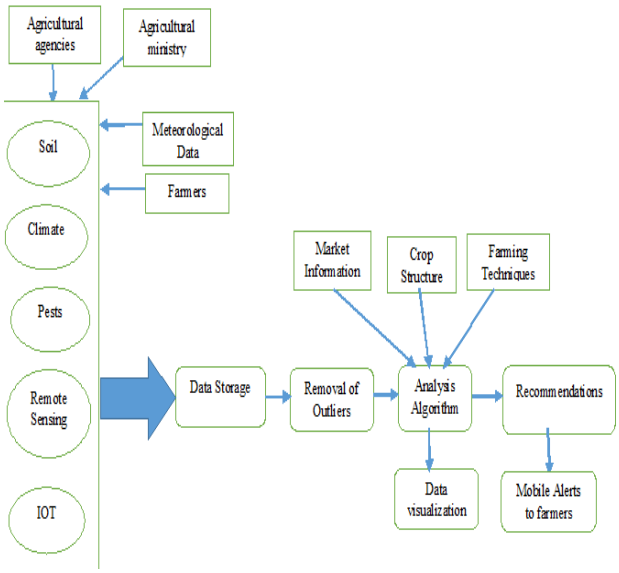


**Figure 3: Big data framework**

The collected data is unstructured and heterogeneous, NoSQL technology has gained popularity to store such data. Agricultural data has to be cleaned by removing outliers and noisy data. A NoSQL data model such as Mongo DB, Couch DB, and HBase is best suited to store the heterogeneous data and to perform preprocessing. Data analysis is the core component of the architecture as it approximates the agricultural issues. Data analysis engages different algorithm and techniques. The open source big data tool such as Mahout is appropriate for implementing a machine learning algorithm over big data platforms. Mobile alerts related to crop selection and best farm practices can be delivered to farmers to adapt to the analytic suggestions. The results generated after the analysis can be best understood with data visualization. With the visualization tools such as R, Rapid miner, D3 it is easy to identify cropping patterns, weather patterns, price variation, etc. In the following subsections, the most relevant sources for big data and their influence over the crop yield are considered along with the techniques used for the analysis of the data.

## V. AGRICULTURAL DATA SOURCES

The crop yield and area of cultivation of all India level is available on monthly and yearly Open Government Data (OGD) Platform India. This data also describes the crop production in differential years and the availability of land for cultivation. The records of the land ownership of the farmers, farming practices followed, soil factors is subject to availability in Department of Agriculture district centers and NBSS and LUP (National Bureau of Soil Survey and Land Usage Planning). The agricultural yield is dependent on geographical (location), Soil conditions, environmental (rainfall, temperature, sunlight), biological (the type of crop, diseases), technical (sensors, irrigation methods) and economic factors.

**Table 1: Factors affecting the crop growth relative to soil**

| Factors | Soil characteristics |
|---|---|
| Temperature, light energy for plant growth | Max and Min Temperature, day length and Sunshine hours. |
| Moisture availability in crop growing season. | Rainfall, PET, Soil depth, and texture. |
| Root development and anchorage. | Root zones Texture, Soil depth, soil structure. |
| Oxygen availability to roots. | Drainage, Moisture retaining capacity of the soil, Levels of groundwater. |
| Nutrient availability. | Organic matter, CEC, pH, NPK status |
| Workability and management. | Slope, Surface stoniness, Moisture holding capacity of the soil. |

### A. Soil

Consideration of soil characteristics and their impact on crop production is a significant element for location-specific crop management systems. The soil properties include physical and chemical properties. Physical properties comprise texture of soil, water holding capacity, rooting depth, structure, and slope [22]. Chemical properties mainly considered are soil total nitrogen (TN), organic carbon (SOC), available phosphorus($P_{ava}$ ), potassium (K), acidity(pH), electrical conductivity (EC), cation exchange capacity(CEC)[7]. Soil properties change over time where physical properties change over 10-15 years and chemical properties change frequently. The changes in the soil properties impact rainwater and nutrient movement within the soil. It also affects water and nutrient supply to plants, growth of roots and crop response [4]. Understanding the soil
properties and their variability is required for effective crop management. The factors affecting the crop growth relative to soil and the soil parameters to measure them is shown in
Table 1. Each crop requires specific soil conditions for optimal yield. In order to rationalize land usability, the soil suitability for various crop yields needs to be analyzed. Although a large amount of data on crop production have been generated by crop research Institutes, it is yet to be correlated with soil-site conditions in order to calibrate soil suitability model to optimize land use contributing to crop productivity. The factors affecting crop growth and the relative soil properties are as shown in Table1:Soil properties and the crop yield are strongly related. Soil quality assessment identifies the soil characters that mirror the capacity of the soil to manage crops [7]. A model relating the favorable soil attributes for the individual crops assists the agriculturists to choose the crop that best suits the soil conditions. But detailed information about soil conditions is limited to certain areas due to the high cost associated with manual soil surveys [34]. Lately, acceptance of the Global Information System (GIS) to study soil attributes is generating accurate soil maps. Contemporary alternatives such as GPS, remote sensing and digital evolution models provide acceptable values of soil properties [45].

The papers mentioned in Table 2 presents the key soil parameters and their effect on yield and the methods associated with the evolution of soil parameters.

**Table 2: Analysis of the role of soil in agriculture**

| No | Reference | Problem Description | Solution |
|---|---|---|---|
| 1. | Sorenson.P.T., et al., 2017 | Examining the Soil Organic Carbon(SOC), Total nitrogen, Soil ph., key soil parameters, | Continuous wavelength spectroscopy combines with machine learning used to measure the soil properties. |
| 2 | De Paul Obade, et al., 2016. | Considering the relationship of soil quality index (SQI) and agricultural productivity. | Nonparametric methods (i) Reduced Regression,(ii) Principal Component Regression, (iii)Partial Least Square Regression(PLSR) evaluated. PLSR is considered suitable for building SQI. |
| 3 | Ayoubi S., et al.,2009 | Correlation of soil variables to explain the variability of barley biomass and grain yields. | Correlated attributes included and arranged in multiple regression models. This correlates biomass and crop yield with soil characters using factor analysis. |
| 4 | Bunemann., et al., 2018 | Soil quality is evaluated with reference to soil functionalities, soil specific threats, and ecosystem services. | Different soil quality indicators are used with different weightings, soil quality assessment and monitoring with technological developments. |

## B. Climate

Climate is the abstract term used to represent the weather conditions- Temperature, rainfall, atmospheric pressure, humidity, wind, sunshine, cloud cover and precipitation [49]. Climate data is *medium volume* as factors considered are fixed, *high velocity* as the decisions are to be taken at real time and *high variety* as the data acquisition is from heterogeneous sources such as weather stations, historic information (weather conditions and climate data, earth observatory), remote sensing ( satellite based). Climate variations are one of the prime reasons for the fluctuations in gross crop production globally. The developing countries rely more on rainfall for agriculture. Climatic factors directly affect water availability. Rain-fed agriculture productivity can be increased with the analysis of the weather patterns and choosing the crops accordingly which contributes to reducing poverty and food insecurity in rain-fed systems [47]. Understanding the effects of climate changes on crop yield requires a model which describes the response of crops to weather changes. This can be perceived with the use of statistical models that are acquainted from historical yield data and measurements of weather such as growing season average temperature and precipitation [24]. The information about growing season characteristics in advance of the season and predicting the climate fluctuations such as variations in temperature and rainfall for a season offers opportunities to improve agricultural risk management. Agricultural yield and food security also rely on water demand and their response to climate and carbon dioxide (CO2) changes considering the fact of the increase in heat exposure [48].

**Table 3: Analysis of the role of climate in agriculture**

| **No** | **Reference** | **Problem Description** | **Solution** |
|---|---|---|---|
| 1 | Tesfaye, K., et al., 2016. | Millions of acres of maize in Africa become affected by drought in the rain fed areas resulting in a 25% decrease in the yield on an average annually. | A Framework that evaluates the performance of new Drought Tolerant crop variations across in south Africa regions by means of geospatial analysis and spatial crop modeling techniques that allows integrated analysis. |
| 2 | Zhang X, Xu M, Sun N et al (2016), | Climate changes and alterations in field management can alter the soil organic carbon and soil nitrogen thus effecting the crop productivity | Variations in fertilizer management and choosing efficient fertilizer strategies to sit the climatic changes. |
| 3 | Lobell D.B, Burke M.B, (2010) | Developing a model to show the crop response to weather in order to forecast the effects of climate on the crop yield. | Using the temporal and spatial statistical models are extensively used to study the consequence of current and future climatic variations on the crop yield. |
| 4 | William W. Guo • Heru Xue (2012) | Factors that are closely associated with the crop yield among different yield affecting factors have to be identified. | Rainfall, temperature, and plantation area are considered and multilayer perceptrons are used to similate the yield. |

A study shows that change in climatic conditions and field management alter soil organic carbon and nitrogen cycling which in turn affects the crop growth and yield [58]. Climate smart agriculture (CSA) is a framework proposed by the World Bank and other international organizations that propose a frame work to embrace the climatic conditions and planning the crops [28]. The influence of temperature, humidity, and sunlight together is best understood using Potential Evapotranspiration (PET) calculations. PET values specify the amount of water evaporated from the surface of the earth due to the above-mentioned factors [31]. PET is calculated using the Thornthwaite equation,

$$PET = 16\left(\frac{L}{12}\right)\left(\frac{N}{30}\right)\left(\frac{10\,T_d}{I}\right)^{\alpha}$$

Where

PET= Potential Evapotranspiration (mm/month)

$T_d$ = average daily temperature of the month is calculated

N= Number of days in a month is calculated.

L = Average day length of the month is calculated.

I= Heat index which depends on the 12 monthly mean temperatures $T_{mi}$

$$I = \sum_{i=1}^{12}\left(\frac{T_{m_i}}{5}\right)^{1.514}$$

PET measurements help in calculating the irrigation requirements and setting the watering schedules for the crop. The techniques or the approaches used in the reviewed papers establish a relationship between climate and crop yield include Machine Learning, statistical analysis, and modeling, cloud platforms, Map Reduce analytics. Table 3 presents the climate parameters and their effect on the crop yield among the papers considered

### C. Diseases and pests

About 40% of worldwide crop yield had diminished due to several diseases. Accurate and timely identification of crop diseases are essential for pest supervision and extenuation of the possible economic and ecological losses. Early detection of the disease is the key factor minimize the loss qualitatively and quantitatively.[29] Accurate data of disease incidence, diseases severity, effects of the disease on the final yield are important for planning crop protection methods. Analysis of disease data combined with plant phenotype facilitates in understanding the dependence of disease occurrence on environmental factors and categorization of the crop disease data.[52] The papers considered for review of plant diseases provides insights into plant disease detection methods and favorable conditions for pest establishment. Disease data is *high volume* as disease data is dependent on environmental factors too, *high velocity* as the recognition of diseases has to be addressed in real time, *high variety* considering multiple pests and diseases. Pest mapping generally begins with the process of documenting where the pest has been observed and linking with environmental factors and also reporting the geographical mapping of the pest records [13]. Current and future changes in the geographical ranges of agricultural pests needs to be related. It aids in perceiving factors of pest risk and increases the chances of providing realistic pest solutions in future [13]. Detecting the disease in early stage guides the measurements that can be taken to prevent and cure diseases. Image analysis as a research field extract information from digital images using computational techniques. The images originate from wide range of sources such as color digital cameras, smartphones or high-quality cameras designed to record finest category of information in the images [29]. The technological advancements is hyper spectral imaging, which records multiple color bands compared to traditional digital imaging. Hyper spectral remote sensing data, which includes the spectral resolution of less than 10nm comprising of diagnosing spectrum information can be used to sense disease possibilities in green vegetation. Developing appropriate spectral index is helpful in detecting leaf diseases in early stages. [15]. Detection of plant leaf diseases is done by image segmentation and later the diseases are classified. Related images of the diseased leaves are considered for training and testing [46]. Crop phenotyping is the successful application to recognize and observe plant diseases. This involves environmental sensors, quality datasets, facilities for data supervision and investigation [42]. Chlorophyll content of healthy and diseased leaves are analyzed to understand the distinct difference between both using hyper spectral image analyses. [26] Generally normalized difference vegetation index (NDVI) is the simple graphical indicator to assess healthy plantation with images. Crop disease management is of critical concern and should be addressed immediately with long term plans for sustaining or increasing crop productivity. Contemporary research suggests that increasing plant diseases resistance allows them to mount a timely defense against the pathogens [19]. Studies speculate the usage of nanoscale materials can effectively retain the nutrients in the plant and serve as long term reliable plants nutrient reservoir. A recent development proposes to increase the host (plant) resistance by incorporating the nanomaterial into agro practices. The plant that is nutritionally stable is resistant to diseases.[39] Table 4 summarizes the methods involved in the detection of the crop diseases and pests in the papers considered for this review.

**Table 4: Analysis of the role of diseases in agriculture**

| No | Reference | Problem Description | Solution |
|---|---|---|---|
| 1 | Liangxiu Han, et al, 2015 | Detection and identification of specific crop diseases to prevent and control the plant diseases efficiently. | Crop diseases are identified based on marker-controlled watershed segmentation of images, super pixel based feature analysis and classification. |
| 2 | Lowe et al, 2017 | Early detection of the plant diseases for early intervention to control the infection, thereby preventing its spread. | High-resolution plant data is collected using hyper spectral imaging facilitating in classifying healthy and diseased plants. |
| 3 | V. Singh, A.K. Misra, 2017 | Automatic detection of plant disease in the early stages of the disease and reducing the labor of farm monitoring. | A genetic algorithm is used for image segmentation technique assisting in plant disease detection and other plant disease classification techniques are surveyed. |
| 4 | Servin A et.al,(2015) | To subdue the crop diseases by making the plants nutritionally stable. | Using nano scale micronutrients considerably increases the crop yield and suppresses the plant diseases. |

### D. Remote Sensing

Remote sensing allows the charting the geographical characteristics of a region without the physical contact with the areas that have to be measured with the images gathered from the satellites. It has several benefits when applied to agriculture varying from crop monitoring and management, fire and flood risk assessment, drought risk assessment. The explosive growth of data from earth observatories and an increase in the degree of diversity and complexity in remote sensing data should be regarded as the big data. Remote sensed data is of *high volume* and *high velocity* [27]. The NASAs Earth Observing System Data and Information System (EOSDIS) archived remote sensing data which would currently exceed 7.5 petabytes every day. Remote sensing is used in agricultural applications from more than a decade. Predominantly Vegetation Indices [VI] is the combination of two or more spectral bands contributing to highlighting a particular vegetation property [8]. VI measurements include Leaf Area Index [LAI], chlorophyll content, and percent green cover. Normalized Difference Vegetation Index (NDVI) assesses whether the observed target is green vegetation or not. VI and NDVI are significantly used in crop mapping and monitoring. Crop yield can be estimated from remote sensing by determining a relationship between statistical yield measures and VI measures of a single day or cumulative of a season [33]. Remote sensing techniques plays a crucial role in crop mapping providing information for number of purposes like estimating grain supplies (crop yield estimation), amassing crop yield statistics, maintaining records of crop rotation, mapping soil factors with productivity, documentation of factors influencing crop stress, valuation of crop loss patterns due to rainstorms and drought, and observing farming activities [41]. Machine learning such as convolution neural networks applied to publicly available remote sensing data is promising effort in agricultural monitoring [57]. Table 5 presents the technologies and methods employed in areas of agriculture using remote sensing. Remote sensing can be employed to potentially estimate the possibility of occurrence of plant diseases and pests in a particular location. Plant diseases and the pest occurrence is evaluated by monitoring leaf, canopy, and field levels [56]. Another approach in remote sensing focuses on Radiation Use Efficiency (RUE), Light Use Efficiency (LUE) to state that total biomass production is directly proportional to total photosynthetic active radiation absorption [25]. Soil mapping is another significant use of remote sensing. Topographic attributes and NDVI provide the information about soil properties, soil moisture that helps in assessing suitable crop and irrigation method to be practiced. India has really made remarkable advancements in remote sensing technologies. Indian Remote Sensing (IRS 1A) satellite launch dates back to 1998. Now, the satellite data is publicly accessible on the Bhuvan Indian Geo-Platform. The following Indian satellites namely: IMS-1 (Hyper spectral), Cartosat, OceanSat, and ResourceSat. The following products are available to download outside of India – NDVI (Normalized Difference Vegetation Index) Global Coverage, CartoDem Version -3R1 for SAARC countries and Climate products for the North Indian Ocean. Launching IRS satellites in India has enabled soil mapping at 1:50000 scale to associate the soil series. The studies with IRS satellites -1A /1B LISS-II had set the trends of rapid advance and appropriateness of remote
sensing application in soil resources study. [nrsc.gov]

**Table 5: Analysis of the role of remote sensing in agriculture**

| Sl No | Reference | Problem Description | Solution |
|---|---|---|---|
| 1. | Bolton, D. K., & Friedl, M. A. (2013) | Predicting soybean and maize harvest in the central USA. | Prediction based on spectral indices resulting from MODIS data and correlating with Cropland Data Layer. |
| 2 | Yuan, L et al (2017) | Estimating the possible occurrence of the crop diseases and the pests in a region. | Combining vegetation indices derived from Worldview 2 and Landsat 8 and environmental factors could precisely reflect the regional spatial distribution of crop and pests supporting early detection of diseases |
| 3 | Mehammednur et al. (2013) | Predicting soil attributes from soil mapping. | Attributes are detected by defining the spectral reflectance of the plants using the satellite images captured during the dry season. |
| 4 | You J, et al (2017) | Real time crop forecasting throughout the year where it is difficult to conduct field surveys. | Dimensionality reduction methods built on histograms and Deep Gaussian Process framework effectively eliminates spatially associated inaccuracies to learn more effective features to predict crop yield. |

### E. Data from IOT related devices

Internet of Things (IoT) is a revolutionary technology that enables the connectivity of object and devices such as sensors, actuators to large databases via the internet. It facilitates interaction between the objects. The data output of different devices is of variable formats. Hence it is crucial to come up with a common protocol for communication between the objects and devices in the network. [30]. Apart from monitoring the crops during the production, the agricultural products have to be tracked after harvest. Using the Wireless sensor networks (WSN) helps in monitoring the storage and logistics facilities of the yield. Radio frequency identifiers (RFID), wireless sensors networks serve as basic building block for using IOT in agriculture. Sensors can be installed on a large scale with low maintenance. Sensors can be deployed to measure environmental factors such soil moisture, soil ph. levels, sunshine, temperature and so on. [9] The data is generated from different kinds of sensors such as scalar sensors to sense the changes, multimedia sensors that capture the images for the detection of plant diseases or monitor the growth patterns and tag-based networks such as RFID to track the products. Unmanned Ariel Vehicles (UAVs) and drones are in use for crop health monitoring, pesticides, and fertilizer spraying.[16] These helps the farmers to understand plant health, plant height measurements, chlorophyll measurements, weed pressure mapping. Enabling actuators for activities such as irrigation and fertilizer diffusion is an efficient way of utilizing resources [55]. Hence different network nodes generate diverse data and the data is semi structured or unstructured. From the discussion, it is evident that IOT witnesses the massive flow of data which needs increased storage and computational resources. [50] It is clear that the success of IOT can be realized by integrating IOT with big data. Figure 4 illustrates the possible framework for IOT and big data where the value is derived out of the data collected from different sources in the network. Automated mechanisms to derive the ubiquitous data from machines into big data is extremely important. Big data analytics examines small and big datasets to extract meaningful values from the data which contributes to making progressive decisions in the farming process. [3]

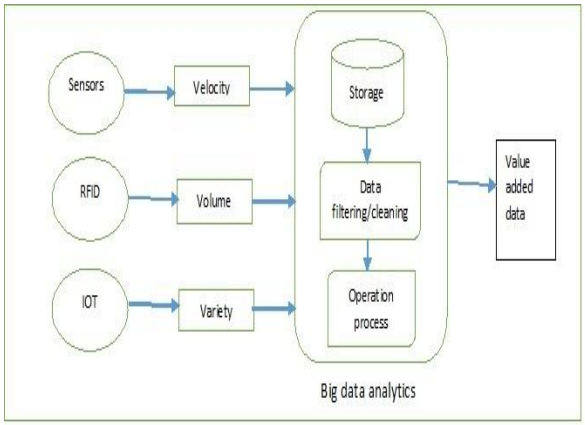


**Fig 4: Integration of IOT and Big Data**

## VI. MACHINE LEARNING FOR BIG DATA ANALYSIS

The value of big data can be realized in decision making with the use of machine learning algorithms. Machine learning is a multidisciplinary field of computer science, artificial intelligence, and statistics. The process of agriculture from seeding to harvesting is biological and involves diverse information to comprehend underlying science and develop decision model. Lately machine learning is being used be the data analysts to leverage the information hidden in big data by determining the associations and understanding the patterns and trends of the data collected [38]. In agriculture, a huge amount of data collected regularly needs to be analyzed and interpreted using machine learning techniques. The data interpretation must reach the farmers with decision support tools. [53] Research scholars are trying to develop a large scale data analytics tool using machine learning opportunity. Agricultural data is heterogeneous from different repositories (statistical, images, unstructured data, etc) could have enormous significance for a learning task. Apache Mahout is a library of machine learning algorithms, it can be implemented over big data platforms. Data collected marks a lot of issues to apply machine learning directly.

Data transformations are necessary to address issues such as data redundancy, noisy data, inconsistency, data imbalance. [59] Big data creates a critical need to develop a novel machine learning solutions to address the challenges of the agricultural data.

## VII. CHALLENGES

India accounts for 159.7million acres of agricultural land, which is second largest in the world after the USA. Agricultural yield is low compared to other countries due to lack of use of machinery and substantial support from the technologies. In developing country like India where most of agricultural land bare small fragments and are family owned. Creation of big corporations in the agri industry and the dependence of the farmers on these monopolies for the farming operations is a necessity from a socio political view. Integrating IOT and big data poses the challenge of data security during the process of data transfer from the wireless sensor medium into data storage. Henceforth IOT architecture is prone to risks such as the man in the middle attacks, illegal access, and the virus injection into the destination. The data collection or the data acquisition needs to be automated to make IOT integration cost effective [17] [44]. Data collected from machines may involve a high proportion of noise due to which the analysis results may vary. Practicing big data applications in developing countries is a challenge due to the imbalance in the availability of technology to all the farmers [17]. With respect to volume and variety, the data diversity may come down due to limited access to technology. The deficient human resources and the experts in the agricultural big data is a major challenge. Hence it is crucial to building a business model which brings together the different stakeholders of agriculture for collecting data evenly from all the regions.[20] Creation of appropriate databases utilizing the public sector and private sector data is essential to tailor cropping patterns, fertilizing methods and use of water resources. Availability of agricultural data is large and requires a protocol for collecting, storing and analyzing the information in order to translate it into decisions. The data needs to be stored for a long time for analysis which necessitates high capacity centralized storage. Transferring of data from fields located in the rural area to the centralized servers requires internet connectivity. [53] In India, internet connectivity is yet to be established in rural areas. The correlation and integration of data from various sources with the different format is the challenge to derive required decisions out of it. A framework to create awareness among the farmers to embrace new technologies and adopt the decisions related to cropping, fertilizers, irrigation needs to be in place.

## VIII. CONCLUSION

Digital data and tools have created the opportunity to gather the data from various sources, directly and indirectly, affecting the crop yield. Big data is a promising platform for storing heterogeneous data. Applying predictive models enables analyzing and managing the data which could be a giant step towards developing an effective decision support system in farming sector. Precision agriculture technologies enable farmers to apply farm specific decisions powered by the use of IoT. Availability of data from different factors considered in this review does contribute to identifying the yield gap in agriculture. Yield gap is the measure of the difference between yield potential of the field and the average yield of the same area [51]. Applying big data analytics aims at reducing the yield gap. Hence promotion of technologies in agriculture anticipates more food production over less land. This paper focused on the review on use of big data in the agricultural sector. The important factors affecting the crop yield are identified and summarized. The accessibility of open big data analysis tools encourages the research towards smart farming. The emerging applications in agriculture need to increase food production and security by enabling the farmers to adopt best practices in agriculture with the suitable recommendations that are environmentally beneficial.

## REFERENCES

1. http://en.wikipedia.org/wiki/Agricultute
2. Abawajy, J. (2015). Comprehensive analysis of big data variety landscape. International journal of parallel, emergent and distributed systems, 30(1), 5-14.
3. Addo-Tenkorang, R., & Helo, P. T. (2016). Big data applications in operations/supply-chain management: A literature review. Computers & Industrial Engineering, 101, 528-543.
4. Ayoubi, S., Khormali, F., & Sahrawat, K. L. (2009). Relationships of barley biomass and grain yields to soil properties within a field in the arid region: Use of factor analysis. Acta Agriculturae Scandinavica Section B–Soil and Plant Science, 59(2), 107-117.
5. Bloom, D.E,2011. 7 billion and counting. Science 333,562-569
6. Bunge J, 2014, Big Data comes to farm, sowing mistrust; seed makers barrel into technology business. Wall street journal
7. Bünemann, E. K., Bongiorno, G., Bai, Z., Creamer, R. E., De Deyn, G., de Goede, R., ... & Pulleman, M. (2018). Soil quality–A critical review. Soil Biology and Biochemistry, 120, 105-125.
8. Bolton, D. K., & Friedl, M. A. (2013). Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agricultural and Forest Meteorology, 173, 74-84.
9. D. Yan-e, "Design of Intelligent Agriculture Management Information System Based on IoT," *2011 Fourth International Conference on Intelligent Computation Technology and Automation*, Shenzhen, Guangdong, 2011, pp. 1045-1049.
10. de Paul Obade, V., & Lal, R. (2016). A standardized soil quality index for diverse field conditions. Science of the total environment, 541, 424-434.
11. Far, S. T., & Rezaei-Moghaddam, K. (2018). Impacts of the precision agricultural technologies in Iran: An analysis experts' perception & their determinants. Information processing in agriculture, 5(1), 173-184.
12. Gartner.com(2013) http://www.gartner.com/it-glossary/big-data
13. Garrett, Karen A. (2013), Big data insights into pest spread, Nature climate change, volume 3, 985–988
14. Gebbers, R., & Adamchuk, V. I. (2010). Precision agriculture and food security. Science, 327(5967), 828-831.
15. Jinbao Jiang, et al, Study on Inversion Models for the Severity of Winter Wheat Stripe Rust Using Hyperspectral Remote Sensing, 2007, IEEE international geo science and remote sensing symposium, 3186-3189
16. Ji-chun Zhao, Jun-feng Zhang, Yu Feng and Jian-xin Guo, "The study and application of the IOT technology in agriculture," 2010 3rd International Conference on Computer Science and Information Technology, Chengdu, 2010, pp. 462-465.
17. Kamilaris, A., Kartakoullis, A., & Prenafeta-Boldú, F. X. (2017). A review on the practice of big data analysis in agriculture. Computers and Electronics in Agriculture, 143, 23-37.
18. K. Kottek, J. Grieser, C. Beck, B. Rudolf, F. Rubel, "World Map of the Köppen-Geiger climate classification

19. Kim, S. H., Qi, D., Ashfield, T., Helm, M., & Innes, R. W. (2016). Using decoys to expand the recognition specificity of a plant disease resistance protein. Science, 351(6274), 684-687.
20. Kshetri N, 2014, "The emerging role of Big Data in key developmental issues; opputunities, challenges and concerns", Big Data and Society
21. Labrinidis A, Jagadish HV (2012), "Challenges and opportunities with big data. Proc VLDB Endowment 5(12):2032–2033
22. Letey, J. O. H. N. (1958). Relationship between soil physical properties and crop production. In Advances in soil science (pp. 277-294). Springer, New York, NY.
23. Liangxiu Han, Muhammed Salman Haleem, Moray Taylor (2015), "A Novel Computer Vision-based Approach to Automatic Detection and Severity Assessment of Crop Diseases", Science and Information conference, July 28-30, 2015, 638-644
24. Lobell D.B, Burke M.B, (2010), On the use of statistical models to predict crop yield responses to climate changes, Agricultural and forest meteorology 150, 1443-1452
25. Lobell, D. B. (2013). The use of satellite data for crop yield gap analysis. Field Crops Research, 143, 56-64.
26. Lowe et al, 2017, Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress, Plant images13:80
27. Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A., & Jie, W. (2015). Remote sensing big data computing: Challenges and opportunities. Future Generation Computer Systems, 51, 47-60.
28. Marcus Taylor (2018) Climate-smart agriculture: what is it good for?, The Journal of Peasant Studies, 45:1, 89-107
29. Mahlein, A. K. (2016). Plant disease detection by imaging sensors–parallels and specific demands for precision agriculture and plant phenotyping. Plant Disease, 100(2), 241-251.
30. M. Lee, J. Hwang and H. Yoe, "Agricultural Production System Based on IoT," 2013 IEEE 16th International Conference on Computational Science and Engineering, Sydney, NSW, 2013, pp. 833-837.
31. Milly, P. C., & Dunne, K. A. (2016). Potential evapotranspiration and continental drying. Nature Climate Change, 6(10), 946.
32. Min Chen,Shiwen Mao,Yunhao Liu, (2014)Big Data: A Survey, Mobile Network Appl 19:171–209
33. Mkhabela, M. S., Bullock, P., Raj, S., Wang, S., & Yang, Y. (2011). Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. Agricultural and Forest Meteorology, 151(3), 385-393.
34. Mehammednur Seid, N., Yitaferu, B., Kibret, K. and Ziadat, F., 2013. Soil-landscape modeling and remote sensing to provide spatial representation of soil *attributes for an Ethiopian watershed. Applied and Environmental Soil Science, 2013.
35. Mondal, P., & Basu, M. (2009). Adoption of precision agriculture technologies in India and in some developing countries: Scope, present status and strategies. Progress in Natural Science, 19(6), 659-666.
36. https://nrsc.gov.in/Agriculture?q=Soil
37. Nandyala Chandra Sukanya, (2016)Big and Meta Data Management for U-Agriculture Mobile Services, International Journal of Software Engineering and Its ApplicationsVol. 10, No. 2 (2016)
38. Garg Raghu, Himanshu Aggarwal, 2016, Big Data Analytics Recommendation Solutions for Crop Disease using Hive and Hadoop Platform, Indian Journal of Science and Technology, Vol 9,32
39. Servin A et.al,(2015), A review of the use of engineered nanomaterials to suppress plant disease and enhance crop yield, springer, Nanopart Res, 17:92
40. Senthilvadivu, S., Kiran, S. V., Devi, S. P., & Manivannan, S. (2016). Big data analysis on geographical segmentations and resource constrained scheduling of production of agricultural commodities for better yield. Procedia Computer Science, 87, 80-85.
41. Shelestov A, Lavreniuk M, Kussul N, Novikov A and Skakun S (2017) Exploring Google Earth Engine Platform for Big Data Processing: Classification of Multi-Temporal Satellite Imagery for Crop Mapping. Front. Earth Sci. 5:17
42. Shakoor Nadia, Lee scott, Mockler C Todd,2017, Current opinion in plant biology, 38,184-192
43. Shah, P., Hiremath, D., & Chaudhary, S. (2016, December). Big data analytics architecture for agro advisory system. In High Performance Computing Workshops (HiPCW), 2016 IEEE 23rd International Conference on (pp. 43-49). IEEE.
44. Sonka, S., & Ifamr, I. (2014). Big data and the ag sector: More than lots of numbers. International Food and Agribusiness Management Review, 17(1), 1-20.
45. Sorenson, P. T., Small, C., Tappert, M. C., Quideau, S. A., Drozdowski, B., Underwood, A., & Janz, A. (2017). Monitoring organic carbon, total nitrogen, and pH for reclaimed soils using field reflectance spectroscopy. Canadian Journal of Soil Science, 97(2), 241-248.
46. V. Singh, A.K. Misra, Detection of Plant Leaf Diseases Using Image Segmentation and Soft Computing Techniques, 2017, Information Processing in Agriculture,4, 41-49
47. Tesfaye, K., et al., 2016. Targeting drought-tolerant maize varieties in southern Africa: a geospatial crop modeling approach using big data. Int. Food Agribusiness Manage.Rev. 19(A), 1–18.
48. Urban, D. W., Sheffield, J. & Lobell, D.B,(2017), Historical effects of CO2 and climate trends on global crop water demand, Nature Climate Change VOL 7 , 901–905
49. Tripathi, A., Tripathi, D. K., Chauhan, D. K., Kumar, N., & Singh, G. S. (2016). Paradigms of climate change impacts on some major food sources of the world: a review on current knowledge and future prospects. Agriculture, Ecosystems & Environment, 216, 356-373.
50. Tzounis, A., Katsoulas, N., Bartzanas, T., & Kittas, C. (2017). Internet of things in agriculture, recent advances and future challenges. Biosystems Engineering, 164, 31-48.
51. Van Ittersum, M.K., Cassman, K.G., Grassini, P., Wolf, J., Tittonell, P., Hochman, Z., 2013. Yield gap analysis with local to global relevance—a review. Field Crops Res. 143, 4–17.
52. Wahabzada, M., Mahlein, A. K., Bauckhage, C., Steiner, U., Oerke, E. C., & Kersting, K. (2016). Plant phenotyping using probabilistic topic models: uncovering the hyperspectral language of plants. Scientific reports, 6, 22482.
53. Weersink, A., Fraser, E., Pannell, D., Duncan, E., & Rotz, S. (2018). Opportunities and challenges for Big Data in agricultural and environmental analysis. Annual Review of Resource Economics, 10, 19-37.
54. Wolfret S, Ge Lan, Big Data in smart Farming- A review , 2017,Agriculrural Systems 153, 69-70
55. Yan-e, D. (2011, March). Design of intelligent agriculture management information system based on IoT. In Intelligent Computation Technology and Automation (ICICTA), 2011 International Conference on (Vol. 1, pp. 1045-1049). IEEE.
56. Yuan, L., Bao, Z., Zhang, H., Zhang, Y., & Liang, X. (2017). Habitat monitoring to evaluate crop disease and pest distributions based on multi-source satellite remote sensing imagery. Optik-International Journal for Light and Electron Optics, 145, 66-73.
57. You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017, February). Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. In AAAI (pp. 4559-4566).
58. Zhang X, Xu M, Sun N et al (2016), Modelling and predicting crop yield, soil carbon and nitrogen stocks under climate change scenarios with fertilizer management in north china plain, Geoderma 265 (2016) 176-186
59. Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. Neurocomputing, 237, 350-361.
60. Kempenaar, C., Lokhorst, C., Bleumer, E. J. B., Veerkamp, R. F., Been, T., van Evert, F. K., ... & van Bekkum, M. (2016). Big Data analysis for smart farming: results of TO2 project in theme food security (Vol. 655). Wageningen University & Research.
61. Waga, D., & Rabah, K. (2014). Environmental conditions' big data management and cloud computing analytics for sustainable agriculture. World Journal of Computer Application and Technology, 2(3), 73-81.
62. Lokers, R., Knapen, R., Janssen, S., van Randen, Y., & Jansen, J. (2016). Analysis of Big Data technologies for use in agro-environmental science. Environmental Modelling & Software, 84, 494-504.


## AUTHORS PROFILE

**Madhuri J** is a research scholar at the Department of Computer Science & Engineering, B.M.S. College of Engineering, Karnataka, India.

She received her B.E. degree in Electronics and communication from Visvesvaraya Technological University, India in the year 2003. She received her Master's degree in Information Technology from Visvesvaraya Technological University in the year 2005. She is currently pursuing her Ph.D. in the field of Big Data and Machine Learning. Her research interests include Big Data, Remote Sensing, Machine Learning and Data Mining.

**Indiramma M** is a Professor and Convener-IIIC at the Department of Computer Science & Engineering, B.M.S. College of Engineering, Karnataka, India. She received her B.E. degree in Computer Science & Engineering from Mysore University, India in the year 1988. She received her Master's degree in Computer Science from Bangalore University in the year 1999. She received her Ph.D. degree from Visvesvaraya Technological University in the year 2010. Her research interests include Big Data, Data Mining, Artificial Intelligence and Data Science