Journal of Theoretical and Applied Information Technology
15th February 2026. Vol.104. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MACHINE LEARNING BASED GCN-LSTM MODEL FOR CROP YIELD PREDICTION USING SPATIAL-TEMPORAL FEATURE LEARNING

MAMTA KUMARI[1], SUMAN[1], DEVENDRA PRASAD[2]

[1]Deenbandhu Chhotu Ram University of Science and Technology, Murthal, India,

[2]PIET Samalkha, India

E-mail: [1]mamta1.tarar@gmail.com , [1]suman.cse@dcrustm.org , [2]devendraacad@gmail.com

ABSTRACT

Prior research has identified limited data and minimal use of soil characteristics as significant shortcomings. To address these issues, this article conducted extensive data collection for bajra yield prediction and introduces a novel Graph convolution neural network and long short-term memory (GCN-LSTM) model, which consists of three main stages: data collection and processing; spatial feature learning; and prediction. The model breaks the limitations of the traditional methods by using data analytics based on IT and advanced deep learning, making more accurate predictions that can be used in smart agriculture, resource optimization, and food security. Unlike previously studied, deep learning models such as recurrent neural networks (RNN), Long Short-Term Memory (LSTM), and Convolutional neural networks (CNN), the proposed GCN-LSTM model does not assume independence among the districts in the crop yield prediction (CYP) data. Instead, it processes attributes related to crop yield prediction such as meteorological, soil, and climate data. The spatial feature learning is not neglected and is leveraged by the LSTM model for the temporal prediction of crop yield. The performance of the GCN-LSTM model is evaluated on RMSE, R^2, and correlational coefficients. The experimental results demonstrate that the proposed model significantly outperforms conventional models by effectively incorporating spatial information.

Keywords: *Graph Convolution Network, Bajra Crop, Soil Data, Spatial Feature Learning, And Soil Characteristics.*

1. INTRODUCTION

By the end of the 21st century, the world's population is estimated to expand to 9.6 billion, significantly increasing the demand for agricultural products. India's population is expected to rise from 1.4 billion to 1.7 billion by the same year, with agriculture covering 1.6 million square kilometers, ranking second only to China [25]. Ensuring food security for its residents is thus a top priority for the government. Modern agriculture practices rely heavily on crop yield prediction (CYP) to allocate resources effectively, mitigate risks, and ensure food security. However, predicting crop yields has become increasingly challenging due to the complex factors involved. Researchers are actively working on methodologies to refine agricultural data collection and prediction techniques [7,18,20]. Despite these endeavors, the diversity in regions, crop varieties, weather patterns, and irrigation methods impedes the development of a uniform and exhaustive dataset for agricultural research [24]. Technological advancements and fluctuations in climate significantly influence agricultural yields, with weather proving to be a notably unpredictable factor that necessitates thorough investigation over time [4]. In this scenario, time- series algorithms, drawing from historical data, have become the preferred approach for predicting crop yields, surpassing other machine learning and deep learning techniques [1-3]. This study specifically targets predicting the yield of bajra (pearl millet), a staple crop in Indian States like Rajasthan, Maharashtra, Haryana, Uttar Pradesh, and Gujarat. Enduring extreme heat, drought, poor soil fertility, and limited precipitation, Bajra thrives where other cereal crops struggle, making it a crucial target for accurate CYP in the state of Rajasthan. This study targets on predicting the yield of Bajra (pearl millet) crops in Rajasthan, India. Bajra is cultivated extensively in regions facing extreme heat, drought, poor soil fertility, and low precipitation, making it resilient to challenging growing conditions. Bajra thrives where other cereal crops struggle, making it a crucial target for accurate Crop Yield Prediction.

The model breaks the limitations of the traditional methods by using data analytics based on IT and advanced deep learning, making more accurate predictions that can be used in smart agriculture, resource optimization, and food security.

1.1 Research Gaps

The limitations identified in the previous studies primarily focus on challenges related to data collection and the narrow scope of attributes analyzed, especially concerning spatial and temporal features. Therefore, the limitations identified in the previous studies include:

- Soil Characteristics were frequently underexamined or included in limited analysis. Neglecting these factors can compromise the accuracy and reliability of crop yield prediction models, as soil quality significantly influences crop growth and productivity.
- Most of the studies utilize convolutional neural networks (CNNs) [12], which typically assume independence among spatial features in the data. However, in agricultural contexts, spatial features often demonstrate complex interdependencies that CNN models may not fully capture.
- Studies such as referenced in [12], explore CNN-RNN architectures, where RNN models are employed for time-series yield prediction. However, RNNs are susceptible to issues like vanishing and exploding gradients. In contrast, research referenced in [4] suggests using LSTM Models for yield prediction in Rajasthan, India, as they are designed to mitigate the limitations of traditional RNNs.

Therefore, it is observed that many studies incorporate time series models such as RNN, LSTM, and CNN to capture the complex temporal correlations (features) among attributes related to crop yield. Moreover, leveraging domain knowledge, particularly regarding crop growth stages, is crucial for enhancing forecast accuracy [26]. However, integrating this prior knowledge into the learning process of temporal correlations remains a challenging task that requires further attention in time series forecasting contexts. Additionally, current deep learning methods [3] [9] [11-12] [17-18] typically assume each prediction unit to be an independent and identically distributed observation, overlooking crucial spatial correlations among neighbouring regions with similar growth patterns in agricultural contexts.

1.2 Key Contribution and Paper Organization

This research aims to address identified limitations by leveraging Graph based Convolutional Neural Networks (GCN) [16] to extract spatial characteristics from data and enhance the accuracy of CYP using the LSTM model. GCN extends traditional CNN to the graph domain, effectively capturing spatial relationships in crop data [17]. By employing GCNs, this study aims to capture the relationships among data attributes and attain higher evaluation performance. Followings are the major contribution of the research:

- Building upon existing literature, this study enhances data collection by integrating a more comprehensive set of variables relevant to crop yield prediction. The study includes a diverse range of factors, covering both meteorological attributes, and soil characteristics, to provide a more robust platform for crop yield forecasting.
- A novel GCN-LSTM-CYP (Graph structured Convolution Neural Network-Long Short-Term Memory architecture - Crop Yield Prediction) system is proposed, structured in three main stages: Data collection and processing; Spatial feature learning; Prediction. This model integrates the neighbourhood knowledge from a constructed graph and temporal insights using LSTM model for CYP.
- Evaluation of the proposed GCN-LSTM model using metrics like RMSE, R2 score, and Pearson correlation. The paper compares its performance against with baselines models including LSTM, RNN, and CNN-RNN.

The paper is organized in the following way : Section 2 covers literature review and related work, and Section 3 outlines the methodology proposed in this study, covering the geographical context, data collection and sources, data preprocessing, and a comprehensive explanation of the GCN-LSTM model. Section 4 outlines the simulation scenario. Section 5 presents the outcomes and analysis. Section 6 offers a comparative evaluation of the proposed model against Reference methods. Finally, Section 7 summarizes the paper with a recap of primary findings and their potential implications.

2. LITERATURE REVIEW

Crop Yield Prediction in Rajasthan state for 10 districts and 7 crops was optimized using Random Forest (RF) and decision trees with gradient boosting. Analyzing data from 1997-2018, the research focused on acreage, productivity, yield and rainfall. Gradient boosting outperformed all other methods, achieving accuracy, with MAE and RMSE Values of 21.58 and 15.01 respectively. Only a few traits and districts of Rajasthan were examined. Gradient boosting is computer intensive and subject to noise, and overfitting can occur with too many decision trees, leading many papers to avoid these methods [1]. Jaipur's agricultural productivity for five crops was forecasted using data spanning from 1991 to 2020, incorporating variables such as precipitation, sunlight duration, temperature and air humidity. Various Machine Learning techniques using Decision tree ensembles and Support Vector classifiers, and Artificial Neural Networks were utilized in the analysis. Random Forest proved to be most precise method, achieving the highest accuracy of 92.3% with MAE and RMSE values of 1.68 and 2.19, respectively. However, it solely focused on one region in Rajasthan and employed a limited set of criteria [2]. In contrast study [3] employs various algorithm's including random forest, SVM, gradient descent LSTM, and Lasso regression. It covers data from 33 districts in Rajasthan across five crops. Results indicate that RF, SVM and Lasso Regression are the top performers in predicting yield. This study acknowledges limitations such as potential for enhancing forecasts by employing a richer and larger dataset for each crop.

In [4], the authors forecast wheat crop productivity across seven Indian states by analyzing satellite photos of wheat fields using and LSTM model. The analysis relies on the Normalized Difference vegetation Index (NDVI) derived from satellite data. Meanwhile, in [5] mustard production in five Rajasthan districts is predicted using long term weather and yield data, Random Forest, SVM, and ANN deep learning techniques are employed. Additionally, Stepwise Multiple Linear Regression (SMLR) and Principal Component Analysis (PCA) are utilized for feature selection, considering factors as climate, rainfall, temperature, soil type and agricultural area. In [6], researchers devised a hybrid model for accurately predicting paddy yield. The model combines Multiple Linear Regression (MLR), with a Feedforward Back Propagation, Artificial Neural Network (ANN), incorporating features like area, number of open wells, tanks, canal length, and maximum temperature during the season.

The researchers in [7] presented a deep learning framework based on Recurrent Neural Networks (RNN) to enhance the Q learning reinforcement learning technique. By incorporating soil, weather and groundwater characteristics, their study achieved 93.7% accuracy rate in predicting paddy output in Vellore while maintaining the integrity of the original data distribution. In study [8], researchers utilized Long Term Time Series (LTTS), alongside weather and soil attributes, Normalized vegetation Index (NDVI), and Supervised machine Learning (SML) algorithms to forecast sugarcane yield in Karnataka. Their methodology involved Support vector regression (SVR) and encompassed the analysis of climatic and soil data, as well as satellite imagery spanning from 2008 to 2018 for remote sensing data acquisition. In study [9], a combination of Artificial Neural Networks (ANN) and Support Vector Machine (SVM) regression models was proposed for predicting the yields of Kharif crops, utilizing rainfall data from Visakhapatnam. Encompassing four crops- Bajra, Rice, Ragi, and Maize- this study utilized a modular artificial neural network to predict monsoon rainfall, followed by support vector regression to estimate the primary kharif crop yields on rainfall and crop area.

An empirical investigation detailed in [10] focused on agricultural production forecasting, particularly regarding the 'Bajra' or pearl millet crop. This study employed regression and time series models/ including ARIMA and ARIMAX with exogenous variables. Results indicated superior performance of ARIMAX over the regression time series model when applied to 'Bajra'. Furthermore, two ensemble models employing Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN) forecasted US Corn production across the corn belt for the period of 1980-2019[11].

County-level meteorological, soil, and management data from 12 US Corn Belt states were collected. Ensembles were constructed using CNNs and deep neural networks, effectively predicting 2019 corn grain yields with an 8.5% relative root mean square error. The study in [12] introduces a deep learning framework for agricultural production forecasting, employing CNNs and RNNs based on environmental factors and agriculture practices. The CNN-RNN model, alongside RF, DFNN, and

LASSO, forecasted corn and soybean harvests in the US Corn Belt using data from 2016 to 2018. The new model excelled over earlier approaches, achieving an RMSE of 9% and 8% for corn and soybean yields, respectively, compared to standard methods. However, the computational demands of ANNs require careful design and precise hyperparameter tuning.

Excessive trees in random forests can lead to overfitting, resulting in computational inefficiency, while Support Vector Regression (SVR) is sensitive to outliers and necessitates careful selection of the kernel function [13]. The emergence of deep learning has seen advancements in complex neural network architectures such as CNNs [14] and RNNs [15].

There is limited research exploring graph-based models to leverage spatial features in crop data for CYP. For example, in [17], the Graph Convolution Network (GCNs) are employed to extract weed plant features from labeled and unlabeled images previously processed by CNNs. The authors focus on enhancing recognition accuracy by learning from graph structures with limited labeled data, although this approach does not directly apply to yield prediction. The survey of dataset attributes used for Crop Yield Prediction (CYP) in recent papers mainly includes meteorological [7], soil [18], phenotype [19], fertilizer, diseases, and pests [20] related attributes, along with other factors depicted in Table 1. These attributes are crucial for determining crop output. In this research paper a consolidated summary of attributes studied in previous research is provided to identify underexplored areas and prioritize further investigation.

Table 1: Summary of Attributes from Previous Studies

Attributes Category (Percentage)	**Attributes**	**Attributes Category (Percentage)**	**Attributes**
Meteorological (40%)	Rainfall	**Phenotype (3%)**	Lodging Percentage
	Temperature		Inversion Percentage
	Humidity		Crop Height
	Sunshine duration		Ear Placement Height
	Evapotranspiration		Barren Stalk Percentage
	dew points temp	**Undesired Traits (14%)**	Weed Determination
	Windspeed		Pest
Fertilizer	Fertilizer		Disease
(7%)	Strength	**Soil (23%)**	Groundwater
Area (3%)	Area under Cultivation		Soil Water
			Soil Moisture
Irrigation accessibilities (3%)	No. of Tanks		Soil Temperatures
	Canal Length		Nutrients
	Open Wells		Soil Ph

Table 1 shows that phenotype characteristics, irrigation accessibility, fertilization, and undesirable features (such as diseases, weeds, and pests) are the least studied attributes with percentage of 3%, 3%,7%, and 14% respectively. These attributes are integral to Precision Agriculture applications leveraging IoT [18]. However, focusing specifically on yield prediction, this study will emphasize other relevant attributes in the prediction process.

Soil characteristics are another key attribute widely utilized in Crop Yield Prediction (CYP). Table 1 also indicates that groundwater levels and soil moisture characteristics are also areas with limited research. However, the current paper incorporates rainfall due to its direct impact on crop yield [21]. Soil types such as red, black, desert, clay, or a mixed soil have received less attention in data collection across studies, making soil type a focal point in this study. Soil type plays a critical role in determining which crops flourish and which struggle; for instance, cotton thrives in dark soil [22]. Furthermore, infertile or barren soils, including with those excessive salt levels, are not considered. This study also examines saline and alkaline soils across various state districts. Additionally local climate conditions are evaluated, as they significantly influence agricultural productivity. Some crops thrive in arid climates, while others prefer humid environments [23]. These deep learning models have become increasingly popular in agricultural production forecasting. According to recent surveys, 69% of studies have adopted deep learning techniques, with CNN-based models being utilized in only 20% of cases.

Comparison of existing work with literature:

- Numerous studies that put greater emphasis on meteorological aspects, including rain and temperature, but simplified or aggregated other soil-related aspects [3, 7, 9,12]. This sparse treatment

can impair predictability because soil quality, including soil type, soil moisture, soil salinity, soil nutrient and groundwater conditions, is a significant factor in crop growth and crop yield [18, 21,23]. Conversely, the current research clearly puts the focus on soils-type differences and related properties as well as climatic conditions and thus covers a major shortcoming of the current research methods and allows the more realistic depiction of the dynamics of agricultural production.

- Recent crop yield prediction works have majorly utilized machine learning and deep learning algorithms like Random Forests, SVMs, CNNs, RNNs, and LSTM-based models [1–5, 9,12]. These methods can be said to be effective in predicting performance but the majority of them assume that spatial units are independent and identically distributed thus missing the spatial correlation present between neighboring agricultural units as similar in soil and climate [3, 9, 1,12]. CNN- and CNN–RNN-based models are efficient at the local patterns and are inefficient at the non-Euclidean spatial dependencies that prevail in agricultural systems [12].
- A number of studies have included the time modeling application with the RNNs and LSTMs to timely ride on time-series dependencies [4,5,7,10]. Nevertheless, RNN-based models have a weakness of vanishing gradient, and despite the fact that LSTMs control such drawbacks, they do not explicitly represent spatial relationships [4]. In addition, most previous studies focus mostly on meteorological factors, and they pay relatively little attention to soil properties, groundwater conditions, and soil type though their influence on crop productivity is also known [18,2,1,23]. Graph-based learning methods have not been applied extensively to crop yield prediction, and all those applications that have been performed so far are centered on plant or weed recognition instead of yield prediction [17].

Proposed GCN-LSTM framework, in turn, directly incorporates both spatial and temporal dependence in the model of districts as graph nodes and acquires to learn inter-regional connections in addition to temporal dynamics. This helps in filling one of the main research gaps revealed in the recent literature [3, 9,1,12].

3. PROPOSED METHODOLOGY

In this section, we discuss the geographical area of study, data collection methods, and pre-processing steps, followed by a detailed discussion on the GCN-LSTM model. The GCN-LSTM model operates concurrently, using GCN for spatial feature learning to extract insights from static features within the CYP data. These insights are then integrated into the LSTM component. We further elaborate on the model's operational details and effectiveness in this process.

3.1 Geographical area under consideration

In India, Rajasthan leads in Bajra cultivation, followed by significant production in Maharashtra, Gujarat, Haryana, and Uttar Pradesh [10]. The challenging weather conditions in Rajasthan limit extensive cultivation of many crops, particularly cereals. However, due to its resilience to adverse climates, millet cultivation has expanded significantly across the region. Rajasthan alone covers for 56.23% (4.43 million hectares) of the total Bajra cultivation region and contributes 41.40% (3.80 million tons) to nation's overall Bajra output [10], fulfilling the state's food requirements.

Figure 1. Geographical area of study

This paper examines the harvest variability of Bajra and predicting its output in Rajasthan's 32 districts as depicted in Figure 1.

This paper focuses on collecting data on soil and weather conditions that have affected Bajra crops in Rajasthan over the years, aiming to predict yield effectively. We aim to capture the yield variability of Bajra and predict it across the 32 districts of Rajasthan. Table 2 provides a statistical summary of the attributes examined in this study and depicts rainfall, a climatic/meteorological attribute, alongside the climate type in Rajasthan. Rainfall ranges from 0.82mm to 8734mm, showing significant deviation from the average, as noted in Table 2. The variability explains the transition from humid to semi-humid climates, which subsequently affects the yield of Bajra [30]. Most districts in Rajasthan experience arid and humid climates. The soil salinity levels also vary widely, from 21.64 to 56369493. High salinity levels hinder crops' ability to retain water and stay hydrated. Adverse climatic conditions further degrade soil quality, leading to lower yields. Consequently, farmers resort to pesticides to enhance crop yields. Hence a structured approach is required to gather data from the appropriate sources.

3.2 Data Sources

The data collection process involves sourcing various attributes from official Rajasthan state government websites. Firstly, the yield data is retrieved from the DACNET website of the Ministry of Agriculture and Farmers' Welfare [32]. This dataset includes yield, area, production, year and district name for 32 districts of Rajasthan, spanning 23 years from 1997. For meteorological data, annual rainfall and climate type are collected. Annual rainfall data is obtained from the Department of Water Resources, Rajasthan Government's official website [31]. This data is distributed across multiple PDF files, each containing daily rainfall records for different districts over the years. The data, available in PDF Format, with separate files for each year, was parsed using Python using libraries like PyPDF2 and regex.

Climate data for yield prediction is sourced from [30]. According to this source, Rajasthan's territory is categorized into five climate types: desert, semi-desert, semi-moist, moist, and Mediterranean. Each of the 32 districts was assigned its respective climate classification based on this categorization. This assignment was accomplished using regex functions to match each district with its appropriate climate type.

Table 3. Column name and description in the final data

Column name	Description
District	Names of the various Districts of Rajasthan used in the present study
Year	The year for which the data was collected
Land Area (Hectares)	Total area in hectares of the district under cultivation for that particular year
Output (Tonnes)	Total production in tonnes of Bajra for that particular year
Yield (Metric Tonnes/Hectare)	Amount of production in tonnes per unit hectare of area under cultivation for a particular year
Soil Type 1	One type of soil found in that district of Rajasthan
Soil Type 2	Another type of soil found in that district of Rajasthan
Phosphorous	Level of Phosphorous concentration in the soil for the district
Potassium	Quantity of the Potassium present in the soil for the selected district
Saline soil Extent (Ha)	Total area of Saline soil in hectares
Sodic or Alkaline soil Extent (Ha)	Total area of alkaline soil in hectares
Yearly Rainfall (mm)	Rainfall for the given district in mm each year
Climate type	The climate type of each district of Rajasthan like arid, humid, wet, etc.

District-wise soil characteristics, including sodium, phosphorus, and potassium levels, as well as the extent of problematic soils such as saline and alkaline soils, were obtained from the Rajasthan Agriculture Statistics at a Glance report available on the official Rajasthan Government website [33]. This data was initially presented in image formats, posing a challenge. To overcome this, Microsoft's Optical Character Recognition (OCR) tool was used to extract the data from images and store it in CSV files. Additionally, soil type data was sourced from another reference [34], which provided information on 16 different categories of soil. Each region within Rajasthan was found to have soils falling into two distinct categories. The final dataset consists of meteorological data, soil-related data, and yield data, encompassing 12 attributes for 32 districts over 13 years. The data is summarized in Table 3 and the overall steps are outlined as follows:

- **Yield Data Retrieval (DACNET Website):** Collected attributes include yield, area, production, year, and district

name for 23 years from 1997 [32].

- **Annual Rainfall Data Collection:** Annual rainfall data is obtained from [31]. This data, available in PDF Files, is parsed using PyPDF2 and regex, then incorporated with the yield data, resulting in a dataset covering 13 years from 2007 to 2019 for 32 districts.
- **Climate Type Data Integration:** Climate type data is incorporated into the dataset from step 2, where each district is assigned a climate type according to [30]. Using regex, the climate type is added to the climate type column for each district.
- **Soil Data Integration:** Soil type data from [34] is added district-wise to the dataset. The district-wise levels of Sodium, Phosphorus and Potassium in the soil (categorized as low, medium, and high) are merged from the data obtained from [33]. This data, originally in images format, is extracted using OCR tools and merged with the existing dataset.
- **Final Data Attributes:** The final dataset comprises 12 attributes: annual rainfall, district name, year, yield, area, production, soil type 1, soil type 2, phosphorus, potassium, saline soil, and alkaline soil, as shown in Table 3.

3.3 Data pre-processing

The first step of data pre-processing is to check for null values. Upon loading the dataset, it is confirmed that there are no null values present. The second step is to identify attributes with an object data type of attributes and encode them. As shown in Table 4, a label encoder from the sklearn library is applied to these object type attributes.

Table 4. Data Type of the Attributes in the CYP Dataset

District	object
Year	object
Area (Hectare)	int64
Production (Tonnes)	int64
Yield (Tonnes/hectare)	float64
Soil Type1	object
Soil Type2	object
Phosphorus	object
Pottassium	object
Saline Soil (Ha)	float64
Alkaline Soil (Ha)	float64
Annual Normal Rainfall(mm)	float64
Climate Type	object

Some attributes labeled as “unnamed: 0”, in the dataset, which contained serial numbers, were dropped during processing. Additionally, certain attribute columns such as potassium, climate type, saline and alkaline soil were named as unnamed and subsequently renamed to appropriate names. For instance, “unnamed: 8” representing potassium was renamed using Panda’s library and regex expressions. The attribute “year” contained value in the format “2007-2008, 2008-2009." To standardize these values, a new list of individual years like ["2007”, “2008” ….”2019”] was created and merged with the existing dataset. Following this pre-processing, the initial step involves constructing graphs based on the years. Further details on these processes will be provided in the subsequent section.

3.4 GCN-LSTM Model

Let’s represent the attributes/characteristics/features of each district as $x_{d,t}$ in the final data and the actual crop output as $y_{d,t}$, where d, t represent district and year respectively. Here according to the Table 1 $x_{d,t} = \{x^{w}_{d,t}, x^{s}_{d,t}, x^{l}_{d,t}\}$ where, meteorological attributes (annual rainfall, climate type) are represented as $x^{w}_{d,t}$, soil attributes (soil type 1, soil type 2, saline soil, alkaline soil, phosphorus, potassium) $x^{s}_{d,t}$, and yield related attributes (area, production, yield) as $x^{l}_{d,t}$. From these features, yield (area, production, yield, soil characteristics) and meteorological attributes (annual rainfall) change both spatially and temporally. Whereas soil attributes (like soil type-1, soil type-2), and climate type, remain stable over time district-wise. Deep learning models like (LSTM, RNN, and CNN) may assume independence among the districts even though attributes like soil type, and climate type remain stable over time but district-wise are different, and while processing these models may not fully use the spatial structure of the original data. To address this issue, this study aims to utilize GCN to extract the spatial features over years h_d, which is leveraged by the LSTM model for time series yield prediction, as illustrated in Figure 3. Hence, the task of CYP is formulated as $y_{d,t}$ with $x_{d,t}$ where $t = \Delta t$ and Δt is set to 3 years (representing the training period for the GCN- LSTM model using 3 years of historical data).

Initially the dataset is employed to construct graphs that depict the interrelationships between

Rajasthan's districts (nodes) across various attributes (edges) over multiple years. Each district is represented as a node in the graph, with its respective name added for clarity. To quantify the similarities between districts based on yield related attributes, cosine similarity matrices are computed annually. These matrices illustrate the relational patterns between districts over time, providing insights into the dataset's spatial dynamics. Following the graph construction, the Graph Convolutional Network (GCN) model is trained to generate spatially learned embeddings. The GCN architecture comprises three layers of graph convolutions with Rectified Linear Unit (ReLU) activation functions. It is configured with specific parameters including input features, hidden channels, and output channels. Training utilizes the Adam optimizer to adjust model parameters iteratively. The GCN extracts spatial features inherent in the dataset, facilitating the creation of an embedding layer through convolutions operations. This embedding layer serves as input for predicting temporal crop yields using a Long Short-Term Memory (LSTM) model. The LSTM component includes layers initialized with predefined input and hidden sizes. It processes the input derived from spatial embeddings, through its layers, culminating in a fully connected layer for yield prediction. Graph construction is facilitated by the NetworkX library, while the GCN utilizes the torch-geometric package [27]. The LSTM model leverages the PyTorch framework [27]. Visualization of the graph is accomplished using Matplotlib and Seaborn libraries.

In the proposed model, the initial step involves constructing a yearly graph to represent the spatial information in the data. Formally, this graph is denoted as a tuple $G = (V, E)$, where V represents the set of nodes and E represents the set of edges (relations between the nodes). The goal is to establish the foundation for spatial analysis by representing the relationships among the districts over each year with V being the districts (32) and E being the the relations between district attributes calculated using the cosine similarity method. The E is stored and represented as a symmetric adjacency matrix $A \in \{0,1\}^{N\times N}$, which consist of N districts for the particular year. Here Ai, = 1 if a relation exists (i.e., cosine similarity value > 0.9) and Ai,j = 0 otherwise. The cosine similarity is computed using the sklearn library, where the cosine kernels, calculate similarity as the normalized pairwise dot product among the districts.

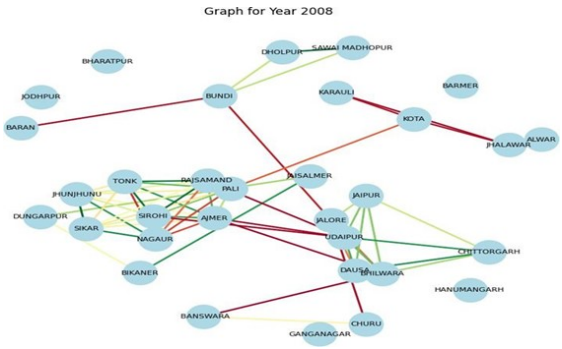

Figure 2. District-wise graph visualization for the year 2008

Graph construction is facilitated by the NetworkX library, while the GCN utilizes the torch-geometric package [27]. The LSTM model leverages the PyTorch framework [27]. Visualization of the graph is accomplished using Matplotlib and Seaborn libraries. Preparing the graph similarity data: Through iterative processing for each year, graphs as shown in Figure 2 are generated using NetworkX, with nodes representing districts and edges indicating similarity based on cosine similarity calculations. Edges are included if the similarity between districts exceeds a threshold of 0.9. The Edge indices are then computed and added to $edge$_indices_list. Subsequently, for each year, a $Data$ object is formed using PyTorch Geometric, containing features representing similarity data for each district and edge indices to denote connections between districts in the graph. This organization of district-wise data over multiple years facilitates the convolutional learning approach adopted from the paper [16]. The graph is visualized for each year. For instance, the graph for the year 2008 is shown in Figure 2. This figure illustrates the dynamic inter-district relationships, showcasing attribute similarity across various districts in Rajasthan as per the map in Figure 2 for each year of data collected. Notably, the absence of an edge connecting Jodhpur and Bharatpur suggests a substantial dissimilarity in attributes explained in Table 2 compared to other districts. This intriguing observation implies distinct characteristics or features that set Jodhpur and Bharatpur apart from their counterparts. The stark dissimilarity, exemplified by the absence of

edges between these districts, underscores the critical nature of certain attributes, potentially hinting at the unique soil types and climate types influencing their yield.

These insights from the graph provide valuable information for understanding regional (spatial) disparities for specific districts exhibiting exceptional attributes in contrast to the broader region, which existing methods often overlook. For each graph, the edges are converted into edge indices resulting in a tensor of shape [2,$umedges$], where $numedges$ is the number of path. Each vertical array in this multi-dimensional array represents an boundary, with the first row containing the indices of the source vertices and the following row containing the destination node indices. This tensor effectively serves as a sparse representation of the adjacency matrix A in equation 3.1. The propagation rule of a l-layer GCN, it is as follows [16],

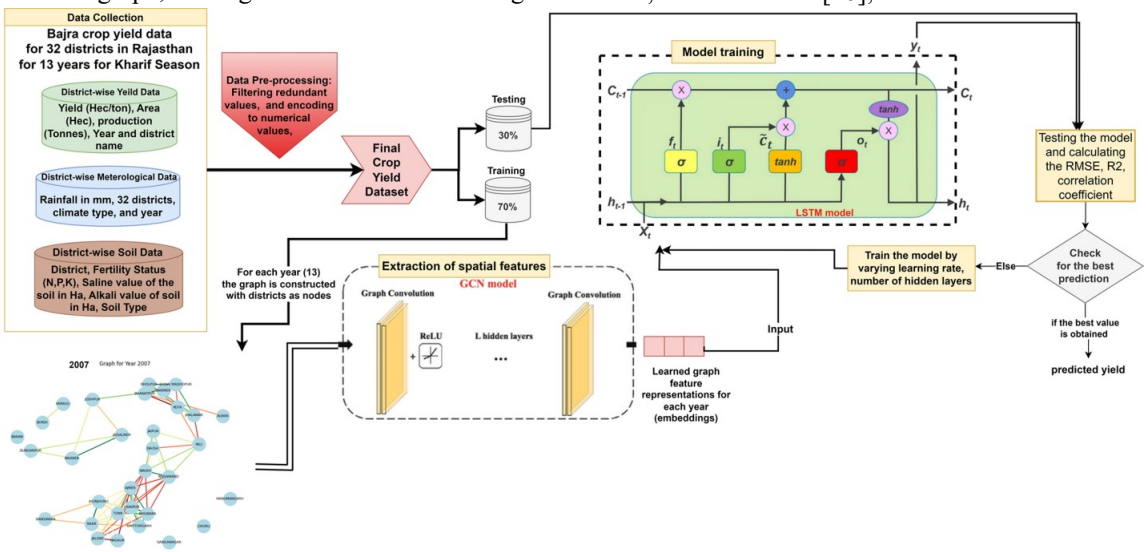

Figure 3. Overall Flow of the Proposed Model

$$H^{(l+1)} = \sigma\left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \quad (3.1)$$

where $H^{(l)}$ represents the node representations at layer l , $A = R + I$ is the adjacency matrix with self-loops added,D is the degree matrix of A , $W^{(l)}$ denotes the learnable weight matrix at layer l , and σ is the activation function, ReLU. Further, the embeddings array generated after processing of GCN hd,t . The final output of the GCN model after passing through all l-layers explained in equation 3.1, is stored in gcn_output. This output represents the node embeddings or features learned by the GCN model for the given input graph. Each row of gcn_output corresponds to the feature representation of a node (district), and each column represents a feature dimension.

4. SIMULATION SCENARIO

The proposed model is implemented on the Google Colab platform using Python and its libraries. The parameters set for GCN and LSTM are given in Table 5, where GCN provides the learned spatial embeddings for span of 13 years for the 32 districts .Graph construction utilizes the NetworkX library, the GCN employs torch-geometric package and the LSTM model is built using PyTorch. Training and testing are performed on the proposed GCN-LSTM model, GCN-LSTM (without soil data) and baseline models: LSTM [3], RNN [29], and CNN-RNN [12]. The GCN is trained using a cross-entropy loss function [16], followed by LSTM for final predictions. Configurable parameters such as learning rate and number of hidden levels are tuned to optimize RMSE, R^2 score, and correlation coefficient, calculated using the Sklearn and SciPy libraries. The study compares several baseline models including LSTM, RNN, and CNN-RNN. The baseline LSTM follows the specifications in Ta, with 70% of the data designated for training and 30% for evaluation. The baseline RNN model [29] uses a window size of 3 and a hidden size of 64, mirroring the baseline LSTM model [3]. Both LSTM and RNN models are trained with the Adam optimizer at a learning rate of 0.001% over 60 epochs to optimize performance.

The CNN-RNN architecture adopted from[12], integrates two CNN models: W-CNN for weather and S-CNN for soil data, each comprising four convolutional layers with specific configurations. In [12] weather and soil data, making up 70% of the training set, are leveraged to model the time based influences of atmospheric variables and the geographical dependencies of soil relative to Bajra crop yield. An RNN component then models the temporal changes in yield over time, similar to the RNN model. Additionally, the research involves training and evaluating the proposed GCN-LSTM model, both with and without soil data, to assess its performance.

In the training process of the GCN model:

- An empty list '$data$' is initialized to store the input data for each year. The list is iterated over the years, appending the similarity data and edge indices.
- The input data for the first year is converted into a PyTorch tensor. The GCN model architecture includes three $CNconv$ layers with ReLU activation functions, followed by dropout regularization.
- The model is initialized with 32 hidden channels and outputs a single channel. The Adam optimizer with a learning rate of 0.01 and weight regularization set to 1e-4 is used.
- During training, the input data is sequentially passed through the GCN layers Each layer applies the learned graph convolution operation anda rectified linear unit (ReLU) activation function to introduce non-linearity.
- Dropout regularization is applied after each layer to prevent overfitting. Optimization of the model parameters is performed using the Adam optimizer to reduce the loss function.
- The output of the GCN model is a set of learned spatial embeddings for all the years as an array. These embeddings capture the structural information of the graph and are used as input for predicting the yield using the LSTM model.

LSTM-based yield prediction process: An LSTM model for prediction is defined, with input size matching the dimensionality of the GCN embeddings, hidden size set to 64, and output size set to 32 to predict yields for 32 districts. The model framework features an LSTM layer which is then connected to a fully connected layer.

- The training loop begins by iterating over a specified number of epochs. Within each epoch, the GCN outputs are fed into the LSTM model in a rolling window fashion as given parameter value in Table 5.
- The mean squared error (MSE) loss is computed between the predicted yields and the true values for each window.
- The loss is then backpropagated through the LSTM model, and the optimizer updates the model parameters to minimize the loss. This trained model is then tested on the test data to get the evaluation metrics.

Table 5. Simulation parameters for proposed GCN-LSTM model

Name of the Model	Parameters	Value
GCN(GCN Conv)	Number of Hidden Channels	32
	Dropout Probability	0.3
	Weight Decay	1.00E-04
	Learning Rate	0.01
	Number of epochs	500
	Input Size	32
	Hidden Size	64
LSTM Model (torch.nn.modules.rnn.LSTM)	Output Size	32
	Learning Rate	0.001
	Number of epochs	60
	Window Size	3

Whereas the RMSE, R^2 and correlation coefficients are detailed as, RMSE is a frequently adopted metric to evaluate the performance of regression models, capturing the deviation between predicted output and real observations by calculating the square root of the average of squared differences between them. Mathematically, RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4.1)$$

Where n refers to the total number of observations (70% samples for training and 30% for testing (10

Journal of Theoretical and Applied Information Technology
15th February 2026. Vol.104. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

years, 3 years)), y_i is the true value and $\hat{y}_i$ is the estimated value. A lower RMSE indicates better performance of the proposed model as it represents smaller prediction errors.

R^2 (Coefficient of Determination) serves as another metric employed to gauge the fit quality of a regression model. It indicates the fraction of variance in the dependent variable (yield) that is accounted for by the independent variables (features). R^2ranges from 0 to 1, where 1 indicates a perfect prediction. Mathematically, $R^2$is calculated as,

$$R^2 = 1 - \frac{SSRes}{SSTot} \quad (4.2)$$

Where, $SSRes$ refers to the residual sum of squares, and $SSTot$ refers to the aggregate sum of squares. R^2helps to understand how accurately the model captures the variation within the data. Residuals are the deviations between the true (actual) yields and the estimated yields produced by the regression model (in this case, the GCN-LSTM model), and on the other hand, the total sum of squares represents the total variability in the observed yields. A higher R^2 value signifies that a greater proportion of the variability in yields is explained by the model, indicating a better alignment of the model within the data.

The correlation coefficient, often denoted as $pearsonr$, measures the strength and direction of the linear relationship between two variables. In the context of yield prediction, the correlation coefficient quantifies how well the predicted yields from the proposed model align with the actual observed yields. Mathematically, it is given as,

$$pearsonr = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}} \quad (4.3)$$

Where, x_i and y_i denote the individual observations of the two variables, $\bar{x}$ $\bar{y}$ indicate their respective means. In this study these metrics are calculated for training as well as testing for all the four models. During training, the model adjusts its parameters to minimize the prediction error (loss) between the predicted and true Bajra yield values. Hence, the model is monitored over each epoch by calculating the evaluation metrics.

5. RESULTS AND ANALYSIS

In this paper comparison is made to evaluate how the proposed GCN-LSTM model performs with the full dataset versus a version that excludes soil data. Both the baseline models and the proposed models were trained and tested on Bajra yield data. The performance was assessed using standardized metrics to evaluate the efficacy of the proposed model. This approach aims to thoroughly assess the effectiveness of our model in predicting crop yield, thereby enhancing the understanding of agricultural forecasting techniques.

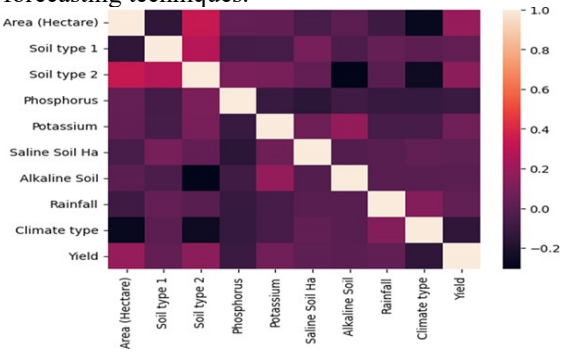

Figure 4. Heatmap representation of the correlation matrix of the attributes

To assess the impact of different attributes on Bajra yield, Pearson's correlation analysis was performed, with results visualized in Figure 4. Data was first encoded using the Sklearn label encoder, and Pearson correlation metrics were calculated using the same library. The heatmap, with a color legend ranging from -1 to 1, illustrates the correlation between various attributes and Bajra Yield. Notably, soil attributes, such as soil type 2, show a strong positive correlation with yield, while climate attributes like climate type exhibit an inverse correlation. This highlights the substantial role that soil attributes play in yield prediction. These results underscore the importance of integrating detailed soil information, which was previously underutilized, to improve the accuracy and effectiveness of crop yield prediction models.

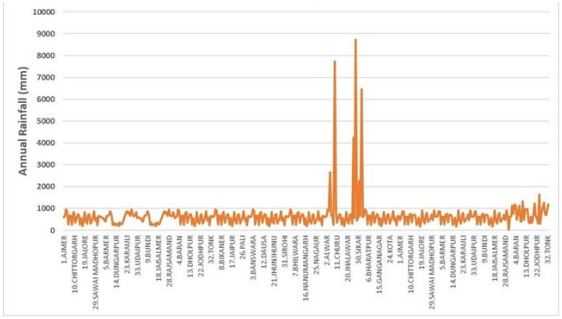

Figure 5. Annual Average Normal Rainfall (mm) over the districts from 2007 to 2020

The time series analysis of annual rainfall, shown in Figure 5, identifies consistent trends across various districts with notable regional variations. Districts such as Alwar, Churu, Jhalawar, and Sikar consistently show high rainfall levels from 2007 to 2019. Correlation analysis indicates a moderate

relationship between rainfall and soil type 2 significantly impacting yield. Based on these findings, the data was trained, tested, and evaluated using the proposed model and baseline models.

5.1 Comparative Analysis

In this section, Table 6 provides a detailed comparison of various models for crop yield prediction, including the proposed GCN-LSTM model and baseline models such as GCN-LSTM without soil data, LSTM, RNN and CNN-RNN. The proposed model demonstrates superior performance by achieving the lowest RMSE values of 7.6 for training and 10.8 for testing, and the highest R^2 scores of 84.8% for training and 68.75% for testing. These results indicate its superior accuracy and fit. In contrast, the GCN-LSTM model without soil data shows higher RMSE values (8.1 for training and 11.0 for testing) and lowest R^2 scores (82.8 % for training and 65.5% for testing), highlighting the importance of including soil data for improved performance. When compared to other models, the proposed GCN-LSTM also surpasses LSTM, which has RMSE values of 95.1 for training and 11.21 for testing and R^2 scores of 74.81% for training and 64.53% for testing. The RNN model with RMSE values of 17 for training and 18 for testing, and R^2 scores of 45.1% for training and 55.1 % for testing, as well as the CNN-RNN model, with RMSE values of 19 for training and 21 for testing, and R^2 scores of 48% for training and 59% for testing, show comparatively poorer performance. This numerical evidence underscores the GCN-LSTM model's robustness and reliability for accurate yield prediction, due to its integration of spatial integration of spatial dependencies through the GCN architecture and its ability to handle long term dependencies better than RNNs, which struggle with the vanishing gradient problem.

The proposed model addresses this by utilizing LSTM cells for long term dependencies and GCN for spatial variations, such as differences in soil composition and climate across regions. As shown in Figure 6, this approach results in an 11.70% improvement in performance compared to the baseline RNN, demonstrating that the GCN-LSTM model better captures the variance in Bajra crop yield data. The spatial representations learned by GCN enable the model to identify and exploit subtle correlations between environmental factors and crop yield, leading to more accurate predictions.

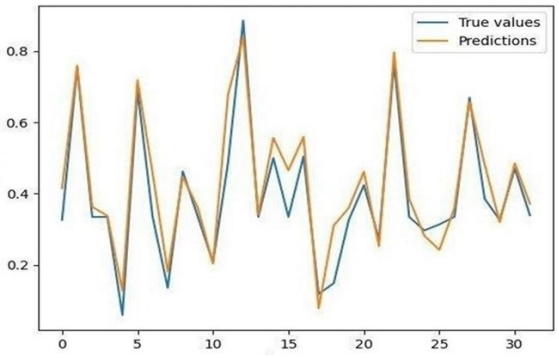

Figure 7. Comparison of True Values versus Predicted Yield Values using GCN-LSTM Model

Table 6. Performance evaluation of Proposed GCN-LSTM Model compared to Baseline Models

Model	RMSE Training Loss	R^2 Training Score	Training Correlation Coefficient	RMSE Testing Loss	R^2 Testing Score	Testing Correlation Coefficient
GCN-LSTM Model (Proposed Model)	7.6	84.8	94.1	10.8	68.75	89.1
GCN-LSTM Model without Soil Data	8.1	82.8	91	11	65.5	83.2
LSTM [3]	95.1	74.81	93.63	11.21	64.53	88.87
RNN [29]	17	45.1	77.1	18	55.1	55
CNN-RNN [12]	19	48	68.6	21	59	50

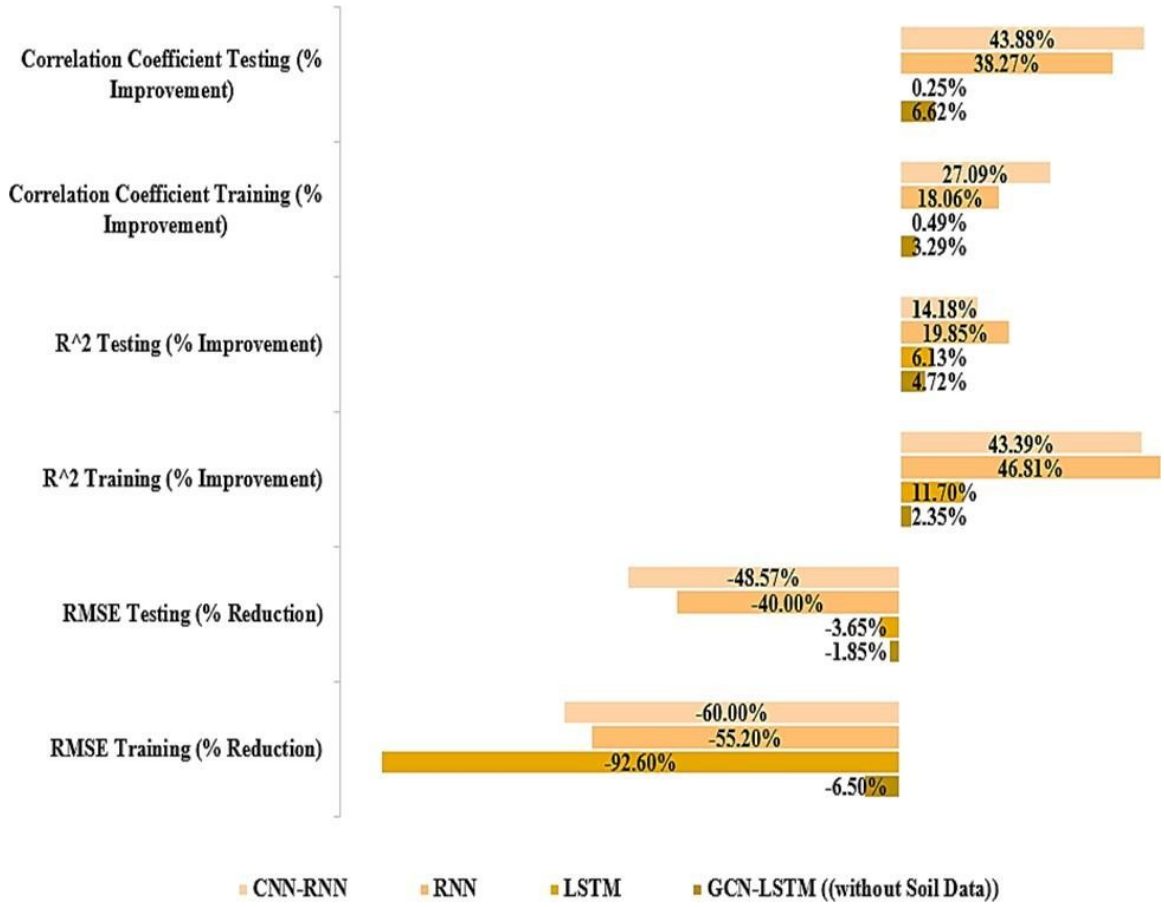

Figure 6. Summary of Performance Analysis for proposed and Baseline models

Figure 7 shows minimal deviation between actual and predicted crop yields, highlighting the model's accuracy. Incorporating soil data, which significantly impacts crop yield, allows the GCN-LSTM model to effectively capture the influence of soil characteristics on Bajra crop yield. These results in improvements in the correlation coefficient during testing of 43.88%, 38.27%, 0.25%, and 6.62% compared to CNN-RNN [12], RNN [29], LSTM [3], and GCN-LSTM (without soil data), respectively. The proposed GCN-LSTM model shows a 3.29% improvement in correlation coefficient during training and a 6.62% improvement during testing compared to the version without soil data. This suggests that incorporating soil data enhances the model's ability to capture the linear relationship between input features and Bajra crop yield, leading to stronger correlations.

Additionally, the LSTM component is adept at modelling temporal dependencies within the time series data. By analyzing historical trends and patterns in Bajra crop yield over time, the model can make more accurate predictions for future yield outcomes. Consequently, the proposed model as shown in Figure 6, demonstrates a reduction of 92.60% in RMSE compared to baseline LSTM and a reduction of 55.20% compared to the baseline RNN model.

5.2 Findings and Discussions

One significant negative aspect of the past literature is the lack of or generalization of soil properties. Some of these studies have focused more on the meteorological parameters, which include rainfall and temperature and simplified/aggregated other factors in soil [3, 7, 9, 12]. This narrow attention may jeopardize predictability because the quality of soil such as soil type, soil moisture, soil salinity, soil nutrients and groundwater conditions are a key determinant of crop growth and yield [18, 21, 23]. On the one hand, the existing study comprehensively considers the differences in the soil types and characteristics, as well as the climatic conditions, which is a significant gap in the literature of the given methods and permits a more realistic representation of the production process in agriculture.

The GCN-LSTM model explicitly learns spatial dependencies, which is one of its major benefits over traditional LSTM, RNN, and CNN-RNN models, which implicitly apply the assumption of independent attributes to their time-series-only models. As compared to the past research which did not consider soil characteristics or the spatial relations, our model shows that the productiveness of the interaction between spatial and time characteristics positively influences the prediction. This gives a clear improvement over traditional models, as the GCN-LSTM version based on the use of soil data demonstrates the lowest RMSE value (7.6 during training and 10.8 during testing) and the largest R 2 value (84.8% during training and 68.75% during testing). The consideration of soil data goes a long way in enhancing the model that forms linear relationships between agro-environmental factors and the Bajra crop yield, as demonstrated by the increase in correlation coefficients of 3.29% (training) and 6.62% (testing) compared to GCN-LSTM that does not include the consideration of soil data. GCN enables the effective representation of space that enables the correlation between districts and LSTM takes care of the temporal long-term dependencies.

Nevertheless, there are still certain shortcomings: the existing model operates with a static spatial graph, and it may not be capable of dynamic inputs of irrigation, land use, or climate. The research only looks at the Bajra crops of Rajasthan therefore making the research not generalizable to other crops and areas. Although the RMSE and correlation measures have improved significantly, further predictive capability of the model can be enhanced by including more agro-environmental variables like temperature, water, or multi-source datasets, with regards to this, potential areas of extension should include: Expand the model to other crops and other agro-climatic areas and enhance its predictive performance; Create dynamic and adaptive learning of graphs, which would enable spatial relationships to change with time as the irrigation, land use, or climate patterns change; Add more agro-environmental features (e.g., water levels, temperature, precipitation patterns, etc.) and multi-source data to enhance the accuracy of predictions. Evaluate the model over long periods to determine the viability of the model in time variation of the data distribution.

6. CONCLUSION

The integration of GCN and the LSTM architecture in the presented GCN-LSTM Model offers a robust solution to the limitations of conventional LSTM and RNN models. The explicit feature of incorporating spatial relations among districts deals with an important shortcoming of the traditional time-based models, the implicit assumption of the independence of attributes. The resulting RMSE and correlation coefficient reductions in comparison to baseline LSTM and CNN-RNN models indicate that the interactions between the spatial and temporal patterns between soil, weather, and crop yield are synergistic because of the intricate interactions between soil, weather, and crop yield. Though the model provides valuable insights into the spatial and temporal dependencies influencing bajra crop yield prediction; the scope is still there for further exploration. Future work may focus on identifying and incorporating additional attributes, such as water levels and temperature, for crop yield prediction. Additionally, an expanded dataset may be gathered to ensure better performance by the model. In this way researchers may refine and extend the exploring the complex interconnections between elements affecting Bajra crop yield; thereby contributing to the improvement of agricultural forecasting methods.

Even though the proposed model is an effective way of capturing both spatial and temporal relationships in the prediction of the Bajra crop yield, and it works better than the conventional LSTM-based and CNN-RNN-based models, there are still some limitations. The present research is limited to the Bajra production in a particular geographical area and this might not be generalizable to other crops and agro-climatic areas as well. In a bid to solve this, the paper will involve extending the framework to include various crops and regions in the future to make it more applicable in various agricultural environments. However, this model uses a static graph design to depict spatial relationship among districts which fails to consider changes in the real world as the irrigation facilities changed, land-use patterns changed, or the climate changed the interactions in the region. Future research will seek to increase the time horizon and test the model on unexplored future years, which is more robust to changing the distribution of the data. The future research will thus focus on dynamic and adaptive graph learning processes in which spatial dependencies will change with time. Furthermore, the paper will focus more towards incorporation of more agro-environmental features and multi sources to be more representative of multi facet nature of crop growth dynamics.

Declaration of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships could have appeared to influence the work reported in this paper. Mamta Kumari, Deenbandhu Chhotu Ram University of Science and Technology, Murthal, India
Suman, Deenbandhu Chhotu Ram University of Science and Technology, Murthal, India
DevendraPrasad, PIET Samalkha

REFERENCES

[1] Jhajharia, K., & Mathur, P. (2024). Machine Learning Based Crop Yield Prediction Model in Rajasthan Region of India. *Iraqi Journal of Science*, 390-400.
[2] SHARMA, S. K., SHARMA, D. P., & GAUR, K. (2023). Machine Learning Techniques for Crop Yield Forecasting in Semi-Arid (3A) Zone, Rajasthan (India). *Current Agriculture Research Journal*, *11*(3).
[3] Jhajharia, K., Mathur, P., Jain, S., & Nijhawan, S. (2023). Crop yield prediction using machine learning and deep learning techniques. *Procedia Computer Science*, *218*, 406-417.
[4] Sharma, S., Rai, S., & Krishnan, N. C. (2020). Wheat crop yield prediction using deep LSTM model. *arXiv preprint arXiv:2011.01498*.
[5] Vashisth, A., & Goyal, A. (2023). Prediction of mustard yield using different machine learning techniques: a case study of Rajasthan, India. *International Journal of Biometeorology*, *67*(3), 539-551.
[6] PS, M. G. (2019). Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms. *Applied Artificial Intelligence*, *33*(7), 621-642.
[7] Elavarasan, D., & Vincent, P. D. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. *IEEE access*, *8*, 86886-86901.
[8] Medar, R. A., Rajpurohit, V. S., & Ambekar, A. M. (2019). Sugarcane crop yield forecasting model using supervised machine learning. *International Journal of Intelligent Systems and Applications*, *11*(8), 11.
[9] Khosla, E., Dharavath, R., & Priya, R. (2020). Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression. *Environment, Development and Sustainability*, *22*, 5687-5708.
[10] Dharmaraja, S., Jain, V., Anjoy, P., & Chandra, H. (2020). Empirical analysis for crop yield forecasting in india. *Agricultural Research*, *9*, 132-138.
[11] Shahhosseini, M., Hu, G., Khaki, S., & Archontoulis, S. V. (2021). Corn yield prediction with ensemble CNN-DNN. *Frontiers in plant science*, *12*, 709008.
[12] Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, *10*, 1750.
[13] Sharma, S. K., Sharma, D. P., & Gaur, K. (2023). CROP YIELD PREDICTIONS AND RECOMMENDATIONS USING RANDOM FOREST REGRESSION IN 3A AGROCLIMATIC ZONE, RAJASTHAN. *Journal of Data Acquisition and Processing*, *38*(2), 1635.
[14] O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.
[15] Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.
[16] Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
[17] Jiang, H., Zhang, C., Qiao, Y., Zhang, Z., Zhang, W., & Song, C. (2020). CNN feature based graph convolutional network for weed and crop recognition in smart farming. *Computers and electronics in agriculture*, *174*, 105450.
[18] Saranya, T., Deisy, C., Sridevi, S., & Anbananthen, K. S. M. (2023). A comparative study of deep learning and Internet of Things for precision agriculture. *Engineering Applications of Artificial Intelligence*, *122*, 106034.
[19] Zhang, Q., Li, B., Zhang, Y., & Wang, S. (2022). Suitability Evaluation of Crop Variety via Graph Neural Network. *Computational Intelligence and Neuroscience*, *2022*.
[20] Sharma, A., Georgi, M., Tregubenko, M., Tselykh, A., & Tselykh, A. (2022). Enabling smart agriculture by implementing artificial intelligence and embedded sensing. *Computers & Industrial Engineering*, *165*, 107936.

[21] Song, X., Wang, P., Yu, J., Liu, X., Liu, J., & Yuan, R. (2011). Relationships between precipitation, soil water and groundwater at Chongling catchment with the typical vegetation cover in the Taihang mountainous region, China. *Environmental Earth Sciences*, *62*, 787-796.

[22] Sudha, T., Ramesh, B., Biradar, D. P., Patil, V. C., Hebsur, N. S., Adiver, S. S., & Geeta, S. (2011). Documentation of cultivation practices of cotton in different soil types. *Karnataka Journal of Agricultural Sciences*, *24*(5), 688-691.

[23] Shi, W., Tao, F., & Zhang, Z. (2013). A review on statistical models for identifying climate contributions to crop yields. *Journal of geographical sciences*, *23*, 567-576.

[24] Gopal, P. M., & Bhargavi, R. (2019). A novel approach for efficient crop yield prediction. *Computers and Electronics in Agriculture*, *165*, 104968.

[25] Visaria, L., & Visaria, P. (1996). Prospective population growth and policy options for India, 1991–2101.

[26] Qiao, M., He, X., Cheng, X., Li, P., Zhao, Q., Zhao, C., & Tian, Z. (2023). KSTAGE: A knowledge-guided spatial-temporal attention graph learning network for crop yield prediction. *Information Sciences*, *619*, 19-37.

[27] Imambi, S., Prakash, K. B., & Kanagachidambaresan, G. R. (2021). PyTorch. *Programming with TensorFlow: Solution for Edge Computing Applications*, 87-104.

[28] McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc."

[29] Bali, N., & Singla, A. (2021). Deep learning based wheat crop yield prediction model in Punjab region of North India. *Applied Artificial Intelligence*, *35*(15), 1304-1328.

[30] Gunawat, A., Dubey, S. K., & Sharma, D. (2016). Development of indices for aridity and temperature changes pattern through GIS mapping for Rajasthan, India. *Climate Change and Environmental Sustainability*, *4*(2), 178-189.

[31] Annual rainfall data date. *Department of water resources, Government of Rajasthan.* Available online: https://water.rajasthan.gov.in/wr/#/department-order/142/23/2776/30900.

[32] Area under cultivation of Bajra.*Area and Production Statistics, Ministry of Agriculture and Farmers Welfare*. Available online: https://aps.dac.gov.in/Home.aspx?ReturnUrl=%2f

[33] Soil characteristics district-wise. *Department of Agriculture, Government of Rajasthan.*Available online: https://agriculture.rajasthan.gov.in/agriculture/#/order/detail/65529

[34] soil type data. Available online: https://agriculture.rajasthan.gov.in/content/agriculture/en/Agriculture-Department-dep/Departmental-Introduction/Agro-Climatic-Zones.html