Sigma J Eng Nat Sci, Vol. 44, No. 2, pp. 1261–1282, April, 2026

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2025.1956

Research Article

Leveraging unexplored regional dynamics and temporal interdependencies in crop yield prediction: A graph theory approach

Mamta KUMARI[1,*], Suman[1], Devendra PRASAD[2]

[1]*Deenbandhu Chhotu Ram University of Science and Technology, Sonepat, 131039, India*
[2]*PIET Samalkha, 132102, India*

ARTICLE INFO

Article history
Received: 29 October 2024
Revised: 23 December 2024
Accepted: 06 March 2025

Keywords:
Crop Yield Prediction; Feature Engineering; Graph Convolutional Network (GCN); LSTM; MixHop; Spatial-Temporal Modelling

ABSTRACT

Agriculture is the pillar of India's economy and food security, yet accurate crop yield prediction remains a persistent challenge due to the complex interplay of environmental, agronomic, and socio-economic factors. In practice, districts often share crop-related characteristics with more than just their immediate neighbours and exhibit irregular temporal patterns in variables such as rainfall. The models that have been implemented frequently do not capture the higher-order spatial relationships between crop-producing areas and also do not take into account the non-uniform timing of the significant agricultural properties, which results in the decreased predictive capability. This paper describes a spatial graph hop with temporal enhancement (SGHTE) framework that utilizes multi-region spatial relationships through a longer MixHop GCN and irregular time dynamics through a time-conscious LSTM. SGHTE is able to capture cross-district interactions, long-range spatial-temporal correlations, and therefore is far more effective at predicting yields in the 32 districts of Rajasthan. The strategy is also enhanced by the use of a fully contained 15-attributes, among which are the underexplored aspects of saline and sodic soil composition, the multiple irrigation sources, the use of hybrid seeds, and the use of fertilizers, which adds depth to the feature space of making the predictions. With a dataset of 32 districts in Rajasthan with 13 years (20072019), SGHTE obtained a 0.1306 RMSE, 0.6775 R 2, and 0.8912 Pearson correlation coefficient to predict the yield of Bajra (Pearl Millet). These findings indicate apparent advances to the state of art approaches, showing that SGHTE has the ability to model intricate spatial-temporal interactions and produce high quality, practical predictions to aid decision-making in the policies of policymakers and farmers alike.

Cite this article as: Kumari M, Suman, Prasad D. Leveraging unexplored regional dynamics and temporal interdependencies in crop yield prediction: A graph theory approach. Sigma J Eng Nat Sci 2026;44(2):1261–1282.

***Corresponding author.**
*E-mail address: mamta1.tarar@gmail.com

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

INTRODUCTION

Globally, approximately 1.3 billion people are engaged in agricultural production, accounting for half of the world's labour force [1]. However, industrialized countries harbour only 9 percent of these agricultural workers and early 60 per cent are in the developing countries [2,3]. The majority of agricultural workers are located in Asia with China having more than 40 per cent of all agricultural people in the world and India having more than 20 per cent [4]. Agriculture continues to be one of the main sources of the livelihood of millions. According to the world development Indicators database, India's population surpassed China's in 2023, reaching 1.417 billion compared to China's 1.412 billion [5]. This population growth raises concerns about future agricultural productivity due to global warming, which causes unpredictable and extreme climate conditions [6]. As indicated by Figure 1, India's population is expected to grow significantly in the coming years [7]. In an effort to address the socioeconomic problem of farmers and satisfy the needs of the fast-growing population, it is important to develop an efficient crop yield prediction model [8]. Nonlinear relationships among different attributes of agriculture reveal more accurate yield prediction with the use of the artificial intelligence (AI) systems to predict the yield [9]. Such systems may be excellent in the provision of proactive measures by policy makers to prevent feminine and mitigate crop failures to farmers [9]. This work is concerned with the forecast of the yield of Bajra (Pearl Millet), a food crop in the arid areas but particularly in Rajasthan, India. The factor that endears Bajra is its ability to withstand extreme climatic conditions in the northwestern and even central India [10]. It grows in semi-arid regions where it is spread over about 7.4 million hectares in India and yields about 9.2 million tonnes yearly [11]. Historically, the prediction of the crop harvest has needed complicated statistical algorithms and models. Nevertheless, the emergence of big data has led to the use of state-of-the-art techniques, such as machine learning and deep learning [12,13]. Such methods as convolutional neural networks (CNN) [14], graph convolutional networks (GCN) [15], recurrent neural networks (RNN) [14], Long short-term memory models [16] have become popular due to their prediction features. The literature review section presents a detailed analysis of past studies that used these techniques. Figure 1 shows the expected population changes of India and China, showing that in 2023, India will overtake China and have 1.4 billion inhabitants. This population change has direct effects on food security and agricultural planning and this has made it necessary to have the right yield prediction models like the one in the present study, hence the urgency to have the right models.

Literature Review

A detailed review of the attributes used for crop yield prediction in Rajasthan is presented, in Figure 2. Relevant Literature on CYP for Rajasthan was sourced from Dimension.ai [17], focusing on publications related to computer science and agricultural production. The identified attributes were categorized into six key groups: environmental, soil, crop, fertilizer, irrigation and seed attributes. According to the reviewed literature [12, 15, 18, 19] and Figure 2, environmental factors such as rainfall, minimum and maximum temperatures, sunlight duration, humidity, wind speed, and production area (in hectares) are the most

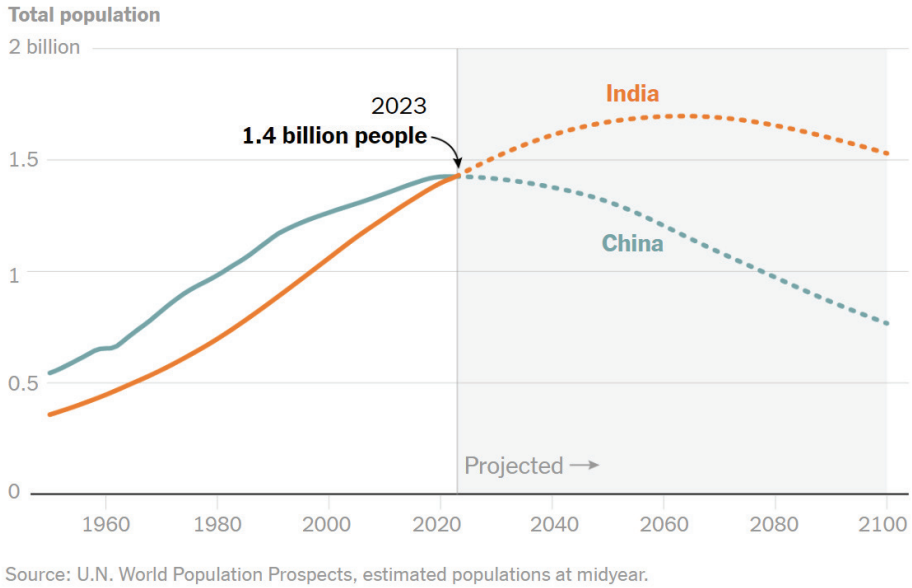

Figure 1. Projected population trends of India and China, highlighting India's expected overtaking of China and its implications for future agricultural demand.

frequently analysed attributes. These factors consistently appear in studies due to their significant impact on crop yield prediction. As Figure 2 also indicates, the agronomically critical characteristics like saline soil content, diversity of irrigation sources, type of hybrid seed and fertilizer are grossly underrepresented in earlier studies underpinning a significant research gap that our study directly fills.

In the research conducted in [12], the authors used the regression technique of Random Forest (RF) to model crop yields in the agroclimatic region of Rajasthan, where six major crops were used. Agrometeorological records of the past were used, such as the records of 1991 to 2020 in terms of rainfall, temperature, humidity, and crop yield. The RF algorithm obtained an accuracy of 92.3%. Likewise, the researchers in [20] used RF and decision trees and gradient-boosting methods to estimate yields of 10 districts and seven crops in Rajasthan. The variables included in the research were acreage, output and rainfall of 1987 to 2018 examined the variables of acreage, output and rainfall of 1987 to 2018. Further, [13] experimented with different machine learning techniques, such as K-Nearest Neighbours, Support Vector Machine, XGBoost regression, and lightGBM that combines the processed data with Smart Farm Ontology to predict yields better. Although they are successful, these machine learning models have a high level of hyperparameter tuning to reach the best performance. They also tend to overfit particularly where the dataset is small or there is a lot of complexity in the model. Moreover, they are vulnerable to noises because of their high-computational intensity, which may compromise the accuracy of predictions. In order to overcome the shortcomings of the classical machine learning methods, deep learning methods, including artificial Neural Network (ANN) were proposed in [21]. In [21] stepwise multiple linear regression (SMLR) was coupled with ANN, support vectors machines (SVM) and RF, and principal component analysis (PCA) coupled with these strategies were used in predicting mustard yield in Rajasthan. On the same note, in the study conducted in [22], ANNs were used together with machine learning algorithms such as RF and logistic regression to predict yield. Although ANNs have enhanced predictive ability, they have high demands on labelled data and extensive computing resources to be trained. This makes them less efficient and more resource-intensive compared to some machine learning techniques, particularly when working with limited datasets or constrained computing environments.

The study in [14] proposed a hybrid CNN-RNN model for predicting crop yields, combining CNNs to capture spatial dependencies in weather and soil data and RNNs to model temporal dependencies in crop yields. Unlike ANNs, CNNs excel at analysing spatial data, effectively detecting critical patterns such as variations in soil quality and weather conditions across different regions. The RNN component addresses temporal relationships by processing sequential data over time. The RNNs however have difficulties with the long term dependencies because of the vanishing and exploding gradient problems. Gradual gradient lead to slow learning of network weights and explosive gradient lead to very large changes in network weights, making the training process unstable. These issues may restrict the capability of the model to learn long time series data.

A CNN-GRU model is proposed in [23] to improve wheat yield estimates using three remotely sensed variables: vegetation temperature condition index (VTCI), leaf area index (LAI), and fraction of photosynthetically active radiation (FPAR). GRUs address the vanishing gradient problem that RNNs lack and facilitate them in capturing long-term dependencies more effectively. GRUs however, do not have a specific state of internal memory cells, which have the ability to store information across longer sequences. This hinders their ability to memorize information over prolonged periods of information. Conventional machine learning models have been used in the prediction of agricultural yield using tabular agronomic and environmental

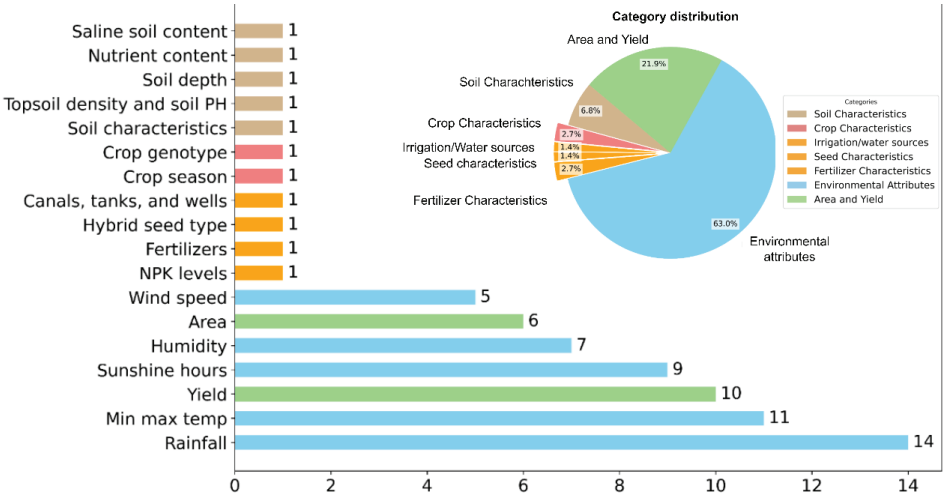

Figure 2. Frequency distribution of attributes for crop yield prediction in Rajasthan, showing environmental factors as most dominant.

data like Adaptive Boosted XGBoost (ABP-XGBoost) [24] and random forest [25]. ABP-XGBoost builds on existing boosting algorithms by dynamically attending to the hard-to-predict samples, whereas the Random Forest combines several decision trees in order to minimize overfitting and increase generalization. These models are also easy to compute, interpret, but do not always have a high-order spatial dependency and temporal anomaly, which can be used as a useful benchmark to measure the performance of more sophisticated spatio-temporal models such as SGHTE.

This study [16] presents a novel deep CNN-LSTM model that estimates soybean yield at the county level in the U.S. In line with previous research [14], the proposed approach employs CNNs for spatial feature extraction. However, for temporal feature learning, LSTM is used. LSTM addresses the constraints of RNNs by integrating memory cells capable of capturing long-term relationships within sequence data. LSTM models are designed to handle sequential data by systematically storing and updating a memory of previous inputs, hence acquiring temporal patterns via their recurrent architecture. However, vanilla LSTMs assume the time interval between events to be regular and uniform, which can lead to suboptimal performance in tasks where time gaps vary significantly [26]. Also, CNNs cannot effectively handle irregular, graph-structured data, and are not effective in capturing both local and global dependencies.

A SCM-GAT (Structural Causal Model Graph Attention Network) model is proposed in [27] that incorporates causal relationships between variables for better interpretability and robustness. GAT utilizes an attention-based mechanism to calculate the connections between nodes in a graph. These attention mechanisms assess the significance of surrounding nodes, enabling the model to concentrate on the most important relations of the nodes [28]. However, GATs can overfit the training data if not regularized, particularly in cases with limited labelled data.

The paper [15] introduces a novel model for accurate crop yield prediction by incorporating GCN to model dynamic spatial topology structures and a knowledge-guided Temporal Multi-head Attention Algorithm (KTMA) for temporal feature extraction. However, traditional GCN has a limited receptive area because, after each convolution layer, it only gathers information from its first-order neighbours.

Challenges to be addressed

Machine learning procedures [12,13] [20-22] and GAT [27] generally have problems with overfitting. CNNs [14,16] have challenges in learning spatial features, such as multifaceted interdependencies in the data. RNNs can experience the vanishing and exploding gradient issues during the processing of the temporal data. GRUs [20] have difficulties with long-term temporal dependencies, and LSTMs [16] fail to deal with irregular time series. Moreover, GCNs [15], GNN-RNN [29] only learn shallow spatial indicators, so they are likely to make inaccurate predictions of the yield, particularly when neighbouring regions follow the same trend of production. The main issues highlighted in the available literature are:

- Some important yield characteristics like saline and sodic soil [13], irrigation sources which are wells, canals and tanks [23], use of hybrid seeds, and use of fertilizers [30] are not discussed anywhere in the literature. These aspects have great impact on both accuracy and strength of prediction models hence poor yield prediction results.
- Environmental factors like soil quality, climatic conditions and irrigation methods bring together crop-producing regions in space [31]. This leads to the fact that such areas tend to exhibit spatial continuity, which yields similar yields [32]. However, most crop yield prediction models fail to capture these spatial connections, resulting in reduced prediction accuracy.
- Many models [14,20,32] assumes that CYP data attributes are uniformly distributed over time. However, several key attributes, such as rainfall and solar irradiation, exhibit temporal irregularities. The irregularity has largely been overlooked in previous search.

The gaps found in the literature, including failure to consider key crop prediction factors, including soil salinity, irrigation availability, and farm fertilizer application, inadequate extraction of spatial relations and time distortions have led to poor models of crop yield prediction. These issues highlight the necessity of more thorough methods. In an attempt to fill these gaps, this paper puts forward a new model that will utilize 15 attributes, including the infrequently explored ones, into the yield prediction model.

Table 2 reflects the gaps in the literature regarding the features of yield, spatial modelling, and the handling of temporal irregularity against the innovations suggested in the proposed SGHTE framework. It is a combination of untested agronomic characteristics, Multi-hop spatial Aggregation, Mixhop GCN, and irregular temporal dependency modelling with t-LSTM to give improved predictive power and robustness at forecasting of Bajra yield in Rajasthan.

Motivation and Contribution of the study

As illustrated in Figure 3, first order neighbors are districts that share a boundary with reference districts. For example, districts such as Sikar, Alwar, Dausa, Tonk, Ajmer and Nagaur are firs- order neighbors but not directly connected to the reference district. In the case of Jaipur, the second-order neighbors are as Jhunjhunu, Churu, Bharatpur, Sawai Madhopur, Bundi and Bhilwara. This rank of neighbors controls the circulation of information among the nodes (in this case, districts), according to their vicinity, and is of paramount importance in the extraction of spatial features. These spatial features that exist between neighboring nodes are extracted well by GCNs. This hierarchical data is important to spatial modelling because most

Table 1. Summary of Research Gaps and Proposed SGHTE Approach for Crop Yield Prediction

Aspect	What is known & research gaps	Proposed approach
Yield-related feature predictors	Previous studies on crop yield prediction in Rajasthan rely heavily on environmental factors—rainfall, min/max temperature, sunlight, humidity, wind speed, and cultivation area [12, 15, 18, 19]. Although these factors are important, there are other agronomic factors (e.g., soil salinity/sodicity, irrigation sources, use of hybrid seeds, use of fertilizers, etc.) that are not well represented, but have a major impact on yield [13, 23, 30]	To enhance predictive power, this research uses a more extensive list of 15 variables of the 32 districts of Rajasthan (e.g. underexplored yet powerful attributes like saline/sodic soil composition [13], accessibility of irrigation through wells, canals, and tanks [23], usage of hybrid seeds, and the application of fertilizers [30]) to enhance the predictive performance.
Spatial modelling (CNN, GCN, GAT, etc.)	CNN-based models [14,16] capture spatial patterns but assume independence between regions, overlooking spatial continuity—where neighboring regions often produce similar yields [32]. Spatial dependencies can be modeled by graph-based methods like GCN [15] and GAT [27]. GNN-RNN [32] captures the apatial connections between the districts but traditional algorithmsonly aggregate first-order neighbor information per layer, limiting their receptive field.	The SGHTE framework builds Mixhop GCN further to combine information about multiple hop neighbours so that wealthier spatial embeddings can be created, which consider local and global relationships between districts in Rajasthan.
Temporal irregularity	Other models such as CNN-LSTM [16], RNN [14]and BiLSTM, GNN-RNN [31] tend to assume time intervals which are uniform yet the data of rainfall and sunshine in the real world show uneven temporal representation. Such a discrepancy is a source of inefficient temporal modeling [32].	t-LSTM is another model in SGHTE that takes into account elapsed time between observations, modifying the memory cell states according to time gaps, and therefore, it exploits the irregular temporal dependencies better.

agricultural properties, including rainfall, soil, and irrigation tend to be spatially continuous beyond direct boundaries. The proposed Mix-hop GCN takes advantage of this Multi-hop connectivity to take into account both the local and the long-range spatial dependency, thus enhancing the accuracy of yield prediction.

As shown in Figure 4, traditional GCNs are limited by their small receptive field, which only allows them to gather information from first-order neighbours through each convolution layer. In this context of yield prediction, it is important to recognize that districts are often connected to others beyond their immediate neighbours [32, 33]. To address this, a Mix hop GCN has been proposed, which enables the model to capture spatial relationships from neighbours that are several hops away. By extracting features from k-different nodes at each convolution layer, as demonstrated in Figure 4[1], this approach improves spatial feature extraction and better captures long range dependencies among districts.

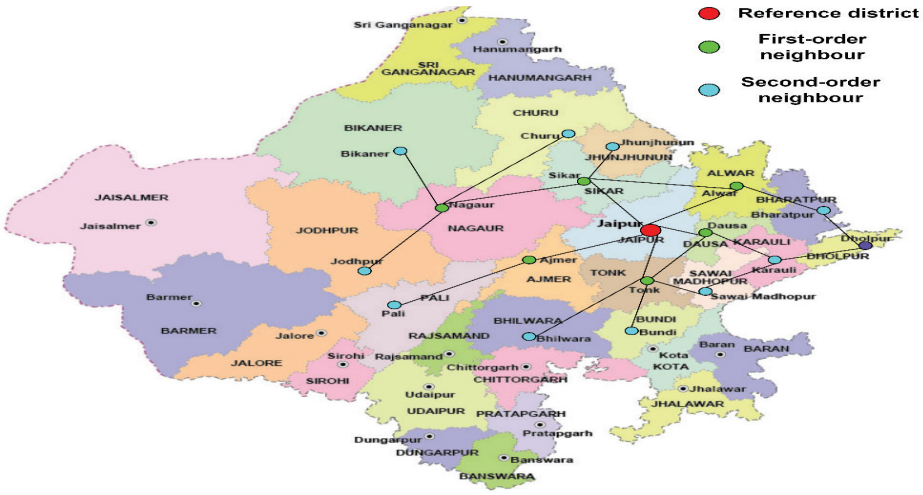

Figure 3. Spatial map of Rajasthan showing reference, first-order, and second-order neighboring districts for multi-hop dependency capture.

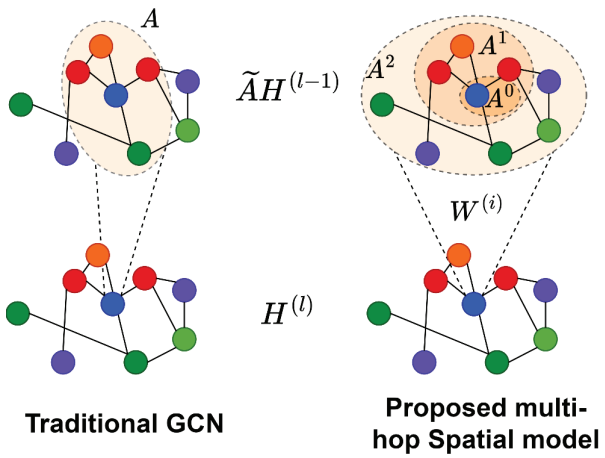

Figure 4. Comparison of feature extraction in traditional GCNs and the proposed Mix-hop GCN, showing how the latter captures higher-order spatial dependencies beyond immediate neighbours for improved yield prediction.

Additionally, it is necessary to note that the crop yield also depends on the spatial and temporal factors, such as precipitation, hours of the sunshine, and annual climatic variations. Figure 5 depicts a rainfall chart of the Ajmer district between the years 2010 and 2020 to emphasize the inconsistency of the data that is provided in the form of the rainfall. Vanilla LSTM Model presupposes that there can be regular time intervals between events [20]. Nonetheless, as it is shown in Figure 5, irregularity in the variability of the rainfall data, a rainfall data, is observed in 2012 and 2014, where uniformity is not observed. A time aware LSTM (t -LSTM) model is developed to overcome this problem of time irregularity in time-related data to extract more temporal features. T-LSTM model is used to provide better time-related feature extraction. T-LSTM integrates a memory, which will adapt the previous cell memory state in accordance with the irregular time intervals between events thus offering a better representation of the time data patterns [33, 34].

To summarise, the main contributions of this paper are:

- The present study presents a new regional data, which will be assembled based on different official sources that include attributes that have not been studied before but are also influential in predicting the yield of Bajra in Rajasthan, including saline and sodic soil composition [13], irrigation availability in the form of wells, canals, and tanks [23], hybrid seeding, and fertilizer application [30]. The fact that these other factors have been incorporated greatly increases the accuracy and strength of yield predictions.
- One of the main contributions of the study is the investigation of multihop neighbouring district relationships based on the graph theory where regional dynamics are used to examine the spatial relationships. To adequately capture the intricate interdependence of geographically related areas, the Mix-hop GCN method is used to make yield predictions more accurate at different locations.
- The given prediction framework can be used to solve time irregularities on the time series with the help of t-LSTM. This is a spatial depth and temporal variation methodology, which is known as Spatial Graph Hop with Temporal Enhancement (SGHTE). Taking into consideration unevenly distributed time intervals, the model is an apt measure of real world time dynamics, which results in higher performance of yield prediction and beats the state-of -the-art models.

Uniqueness of work: Compared to previous models [13,23,30], the paper introduces a regional dataset for Bajra with 15 attributes, including previously underused agronomic factors (saline/sodic soil, irrigation source mix, hybrid seeds, fertiliser usage). The paper focuses on generating multi-neighbour information of districts which contribute to yield prediction, compared to previous models like [15,27,32] which only considered immediate neighbour information of districts. The paper consider d the concept of temporal irregularity which is understudied (motivated

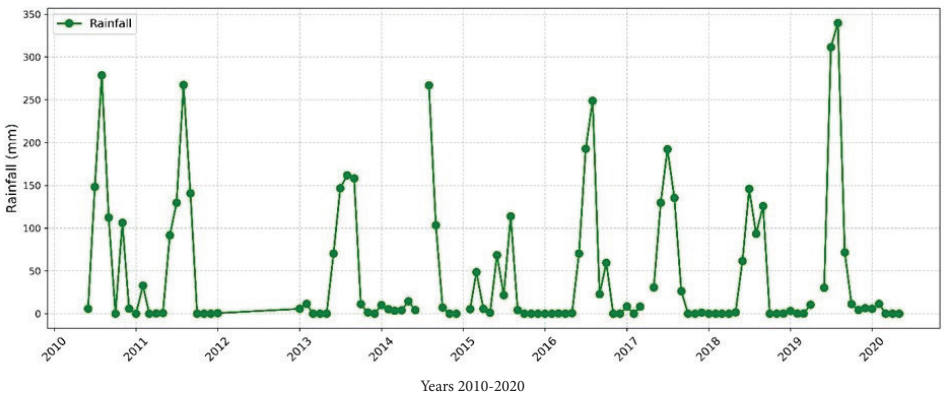

Figure 5. Rainfall variability in Ajmer district (2010–2020), demonstrating irregular temporal patterns that motivate the use of time-aware LSTM for handling non-uniform event intervals in climatic data.

by observed rainfall variability and limitations of vanilla LSTM/RNN) [14,16,21,35].

The main goal of the current research is the development and testing of a new Spatial Graph Hop with Temporal Enhancement (SGHTE) model that incorporates MixHop Graph Convolutional Networks (MixHop GCN) and Time-Aware Long Short-Term Memory (t-LSTM) to increase the prediction of yields in the Bajra crop in the district level in Rajasthan, India. The SGHTE model with its high-order spatial dependencies between districts and its capacity to take into account temporal anomalies in historical yield information should deliver a high level of predictive accuracy, in comparison with contemporary state-of-the-art solutions. The effectiveness of the model is proved by the use of the comprehensive experiments and comparative assessment involving the real-world agricultural data.

The research is an important contribution to the field of agriculture with a solution to a major problem in agriculture, which is the prediction of crop yield, and hence it will fill a gap between the engineering and natural sciences. SGHTE takes advantage of the irregular temporal patterns of the environment, agronomic, and socio-economic factors by integrating Mix-hop GCN to model the spatial dependency of multi-regions and time-sensitive LSTM to model the irregular temporal patterns. Although applied to Bajra yield prediction in Rajasthan, the approach can be applied to other crops, areas, and areas with key spatial-temporal interaction, including climate impact analysis and resource management [35, 36]. The enhanced predictive accuracy has a beneficial practical use by policy makers, farmers and planners in data-driven decision making to achieve food security and sustainable agriculture.

The remaining part of the paper is organized in the following way: In section 2, the proposed methodology is described, including the formulation of the problem, the model architecture, yearly process of the construction of a relational graph of district attributes, and the SGHTE model and the SGHTE model pseudocode. Section 3 explains the design and setup of the experiment, such as data set, preprocessing methods, training and testing procedures, and evaluation measures. Section 4 provides the outcomes of the SGHTE model, and the analysis of the results is done in comparison with the state-of-the-art approaches. Lastly, Section 5 will be the conclusion of the study which summarizes the important findings.

1. Proposed Methodology

This section describes the methodologies that the proposed study will use. The problem formulation section determines and identifies the major issues in crop yield prediction (CYP) and discusses the possible solutions. Section 2.2 presents SGHTE model, which is a combination of Mixhop GCN and t-LSTM to extract spatial feature and learn temporal features respectively. The following section describes how annual district attribute relational graphs are constructed, which is the input of Mixhop GCN model. Lastly, the operational structure of the SGHTE Model is studied further with a focus on how it addresses the issue of spatial temporal dependencies.

1.1. Problem Formulation

The two main unaddressed issues of CYP are the extraction of deeper and more profound spatial relations among the crop-producing regions and the incorporation of temporally uneven attributes. As seen from the literature, existing studies only utilize information from the first-order neighbouring nodes (nodes being districts, cities, counties, etc.) to construct the graphical embeddings ($H^{(l)}$) for yield prediction. Also, the uneven distribution of attributes over a time period has also been ignored in yield prediction. The proposed SGHTE model addresses both of these shortcomings.

Regions with similar latitudes and longitudes have similar climatic characteristics. These patterns show a substantial link between Bajra yields across districts. Indeed, it is seen that a district producing a substantial harvest of Bajra in a given year results in increased yields in the adjacent districts too. Therefore, to consider this spatial relationship, the connections based on regional dynamics by graph theory are incorporated. The data used and the collection procedure for the construction of this graph is explained in detail in section 4. The yearly district attribute relational graphs G_T, which are categorized by districts and provide comprehensive data on many attributes of crops like geographical data (x_{geo}): area, production, districts, year, meteorological data $x_{meteorlogical}$: (Annual rainfall, climate type), $x_{soil\ characteristics}$: (phosphorus, potassium, saline soil, alkaline soil), $x_{soiltype}$= (soil type 1 and soil type 2) $x_{yield\ boosters}$: hybrid seeds, and fertilizer use in Tonnes) and water resources $x_{waterres}$: (wells, canals and tanks) for each particular district on an annual basis. Thus, the set of crop features is represented as $x = \{x_{geo}, x_{meteorlogical}, x_{soil\ characteristics}, x_{soiltype}, x_{yield\ boosters}, x_{waterres}\}$. Thus, the spatial dependencies among crop-producing regions are captured as $f(x, A_t)$. The adjacency matrix A_t captures these attribute relations among x for each year with the help of cosine similarity.

The yearly district attribute relational Graphs are denoted as, $G_{t\in\{2007:2019\}}$ and the information about these graphs is stored as $A_{t\in\{2007:2019\}}$. Thus, to generate the year-wise spatial relations (E_t), the layer-wise (l) propagation is computed as,

$$f(x, A_t) = E_t^{(l+1)} = \sigma\left(\widetilde{D}^{-\frac{1}{2}}\tilde{A}_t\widetilde{D}^{-\frac{1}{2}}E_t^{(l)}W^{(l)}\right) \quad (1)$$

Where, $\widetilde{D} = \sum \tilde{A}_t$ and $W^{(l)}$ is the layer-specific trainable weight matrix, $\sigma(.)$ is the activation function and $E_t^{(l)} \in \mathbb{R}^{N\times e}$ (e is the size of embeddings (year-wise spatial relations)) is denoted as the embeddings in l^{th} layer; $E_t^{(0)} = x$ However, the adjacency matrix is used in the term $\tilde{A}_t E_t^{(l-1)}$ in the vanilla GCN [36, 37], aggregates the features only from immediate neighbours as shown in Figure

3. Given that the convolution operation is carried out only once on the adjacency matrix , the GCN intrinsically prioritizes retrieving information from immediate neighbouring districts in each layer.

In this paper, this is addressed by aggregating information from both first-order, second-order neighbor districts and potentially higher-order spatial relations based on power P [1] as explained in Figure 3. This is enabled by using various powers of the adjacency matrix A in each layer. This facilitates the incorporation of information from k-hop neighbouring districts in a single layer. Thus, the equation (1) is modified as,

$$E_t^{(l+1)} = \|_{j \in P} \ \sigma\left(\tilde{A}^j E_t^{(l)} W_j^l\right) \tag{2}$$

where the hyper-parameter P is a set of integer adjacency powers, and $\tilde{A} = \{\tilde{A}_1, \tilde{A}_2, \cdots \tilde{A}_t\}$ denotes the adjacency matrices with self-connections, as shown in Figure 4 and $\|$ is the column-wise concatenation operator. Thus, the $H^{(l)}$ embedding is the concatenation of all the output year-wise embeddings $E_t^{(l)}$ and is given as

$$H^{(l)} = \odot_{t=1}^{t} E_t \tag{3}$$

To tackle the temporal irregularities as explained in Figure 5, the improved t-LSTM is employed. The input to the t-LSTM network consists of spatially aware node embeddings $H^{(l)}$ obtained from the output of mixhop GCN. Now the hidden state at time t is given by,

$$h^t = tLSTM(H^{(l)}, \Delta t) \tag{4}$$

Here Δt is the elapsed time between two irregular observations. Thus, the final predicted yield $\hat{y}_{d,t}$ is obtained by classification based on irregularity aware hidden state equation (4) and can be given as,

$$\hat{y}_{d,t} = FCN(h^t) \tag{5}$$

Proposed Model

The proposed SGHTE (spatial graph hop with temporal enhancements) model as seen in Figure 6 consists of three primary components: a thorough process of data collection and preprocessing, the Mixhop GCN for generating deep embeddings of the attribute relation of all the 32 districts, and t-LSTM to capture the temporal dependencies of the attributes and predict the yield. Data is parsed from various sources [37,38] The data exists in several modalities, complicating the extraction and collection process. Various pre-processing techniques are applied with the help of various Python libraries like Pandas and Regex to get the final data. This final data set contains 15 attributes regarding Rajasthan's 32 districts. This dataset comprises both temporal and spatial data of the districts. The pre-processed data is used by the SGHTE model. To construct the graphs for the spatial relations, the adjacency matrix is used, which is calculated by the cosine similarity between the edges of the nodes, as seen in step 2 of algorithm 1. The yearly district attribute relational Graphs (G_T) which contains the spatial information and relations are then

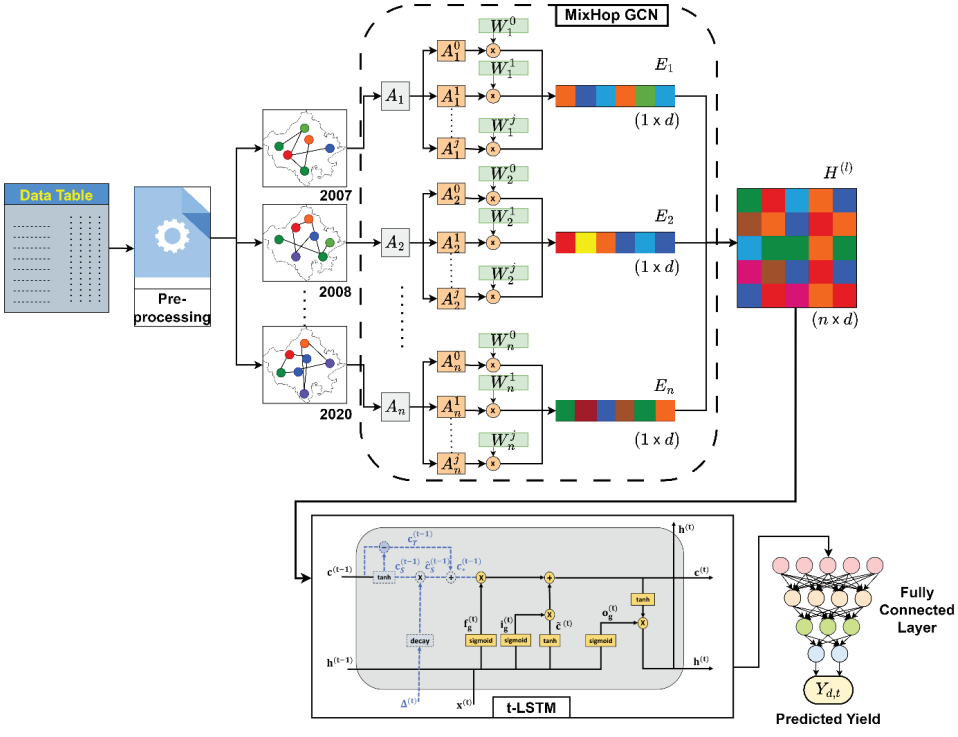

Figure 6. Process flow of the proposed SGHTE model, integrating Mix-hop GCN for spatial feature extraction with time-aware LSTM for modelling irregular temporal dependencies.

constructed using the adjacency matrices A_T (step 3). This G_T is propagated through various layers of Mixhop GCN. Mixhop GCN's convolutional layers gather information not just from neighboring nodes but also from the k distinct nodes around them (step 4). After the generation of these spatial embeddings for each year, these are aggregated in a final embedding $H^{(l)}$ (step 5). These embedding $H^{(l)}$ serve as the input for the model t-LSTM (step 6). The t-LSTM model now processes these embedding $H^{(l)}$ and addresses attributes like rainfall that are not linear over a given time. The SGHTE model is discussed in later sections in detail. The predicted yield is then enhanced by refining the model parameters using evaluation metrics (step 8 and 9).

Construction of Yearly District Attribute Relational Graphs (G_T)

The mixhop GCN model operates on graphs $G_T = (V, \mathcal{E})$ where V is the set of nodes and $\mathcal{E}$ is the set of edges of these nodes. Here Rajasthan's 32 districts serve the role of nodes, while edges represent the connections between them. The district nodes are $V = \{v_1, v_2, v_3 \cdots v_N\}$, whereas the connection among the pair-wise districts (v_i, v_k) is represented by the edge $\varepsilon_{ik} \in \mathcal{E}$. These connections are based on geographic proximity or environmental similarity. The G_T graphs are constructed with the help of the Adjacency matrices A_T. The adjacency matrices A_T represents the connections between the districts in the graph. With total N districts, the adjacency graph is a N x N matrix. A_{ij} contains the weight of the relation between districts i and k which is given by,

Algorithm 1. Overall algorithm

Input: ,
Output: Predicted yield

1. $H^{(1)} = 0$
   For $t = 1$ to t
   For each district $i = 1$ to N
2. Calculate $A_{i,j}$ using equation 6
   End for
3. Construct G_T
   For epoch in 1:100:
4. calculate yearly embeddings $E^{(t)}$ using equation 2
   End for
5. Final spatio-temporal embeddings $H^{(1)} = H^{(1)} \odot E_{(t)}$
   End for
   For epoch in 1:100:
6. Temporal relation mapping $h^t = tLSTM(H^{(l)}, \Delta t)$
   End for
7. Predicted Yield $\hat{y}_{d,t} = FCN(h^t)$
   For $[0,1, \cdots i]$ where $i = \{1, P\}$
8. Optimize *RMSE*, R^2, and r
   End for
   For learning rate= 0.0001 to 0.1
9. Optimize *RMSE*, R^2, and r
   End for
   Final predicted yield $\hat{y}_{d\,t}$

$$A_{i,k} = \frac{x_i^j . x_k^j}{\left|x_i^j\right| . \left|x_k^j\right|} \quad (6)$$

Where x_i^j, x_k^j represents the j^{th} feature matrix of the i^{th} and k^{th} districts in V respectively. Here $j \in [1, M]$ where M are the total number of features and $i, k \in [1, N]$. The following produced adjacency matrices A_T are used for the generation of the yearly district attribute relational graph, denoted as G_T. The aforementioned procedure is repeated annually, obtaining all required $G_T = \{G_1, G_2, \cdots G_t\}$ graphs.

Figure 7 represents one of the yearly district attribute relational graphs, G_9 showing the spatial relationships among various districts in Rajasthan for the year 2016 where each node represents a district, and the edges (lines) between them indicate the strength of the spatial correlation or dependency in various attributes between those districts. Here red and thick edges indicate strong spatial dependencies, suggesting that one district is highly correlated with that in the neighboring districts. While green and thin indicate weaker dependencies or correlations between districts. The graph displays clusters of districts characterized by significant interdependencies, which are accentuated by the red borders. For instance, districts like as Jaipur, Sikar, Ajmer, and Chittorgarh have significant geographical correlations, suggesting that the Bajra yield in these regions tends to follow similar patterns.

SGHTE (Spatial Graph Hop With Temporal Enhancements) Model

The SGHTE model consists of a combination of Mixhop GCN, and t-LSTM architectures as seen in Figure 6, which are used to capture spatial and temporal dimensions, respectively. CNN-based models [14,16] capture spatial patterns but assume independence between regions, overlooking spatial continuity—where neighboring regions often produce similar yields [32, 39]. Graph-based models such as GCN [10] [15] and GAT [27] can capture spatial dependencies. GNN-RNN [32] captures the apatial connections between the districts but traditional algorithmsonly aggregate first-order neighbor information per layer, limiting their receptive field. In order to achieve precise spatial attribute extraction, it is essential to use higher-order attribute capturing. This kind of capturing involves nodes receiving latent representations from their immediate (first-degree) neighbours as well as from subsequent k-degree neighbours at each convolution layer as seen in Figure 6. Considering this the authors use the mixhop GCN, a model that incorporates trainable aggregation parameters to determine the optimal combination of latent information from neighbouring nodes at different distances [40-42]. It is used as a spatial modelling technique to achieve larger and enriched embeddings, and its effectiveness has been shown. The model integrates data from successive K-order neighbours.

The graphs G_T are the input of the mixhop GCN model. They are trained over each year to obtain the spatial embeddings E_n for each year. The mixhop model aggregates

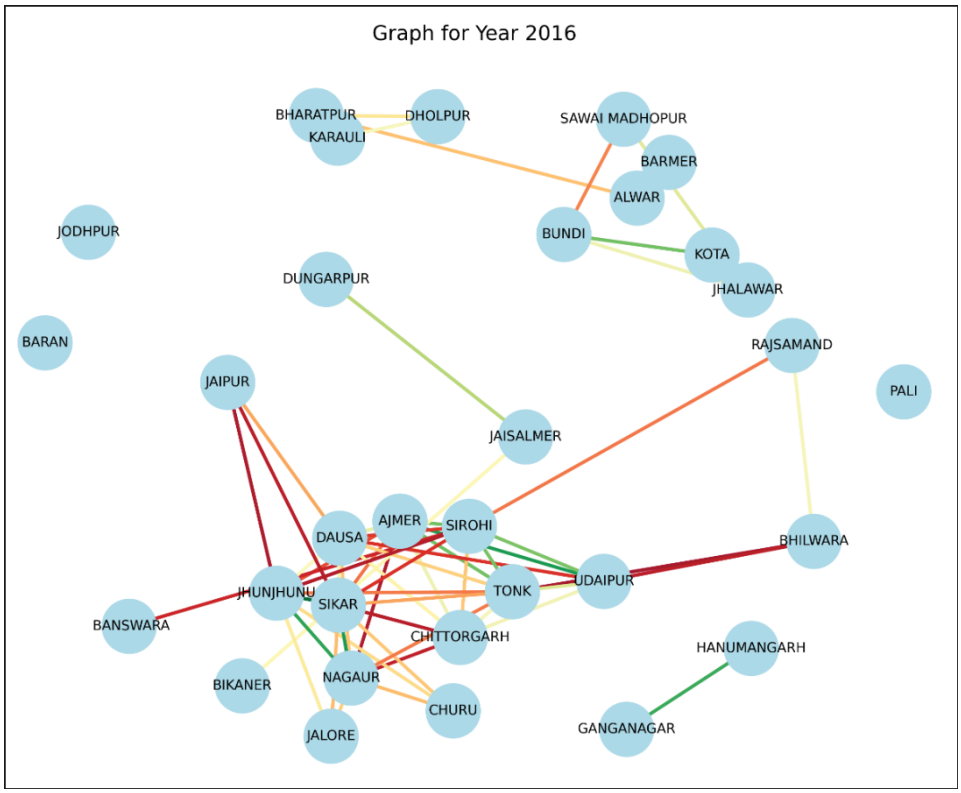

Figure 7. Attribute relation graph of Rajasthan's districts for the year 2016, visualizing spatial connections and feature linkages among neighboring regions.

information from the k-neighbouring nodes by taking the self-product of the adjacency matrix A_T. The number of times the self-product is performed is determined by the hyperparameter P. The adjacency matrix A_T is processed with the help of equation (2). Equation 2 is used to generalize the diffusion convolution layer. The model however is trained using the RMSE (Root Mean Square Error) loss function. The mathematical definition of RMSE is as follows:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_{d,t} - \mathrm{E_t})^2} \quad (7)$$

Here the variable m denotes the number of samples, $y_{d,t}$ represents the actual yield, and E_T represents the projected yield from yearly embeddings. A decreased root mean square error (RMSE) indicates a reduced number of prediction errors, indicating that the proposed model is exhibiting superior performance. The year-wise spatial embeddings derived from Equation (2), are trained using Equation (7). This results in highly feature-rich embedding E_1. Now the model is iterated for all 13 years resulting in $H^{(1)}$ which is the output of the mixhop GCN.

The $H^{(1)}$ embeddings are obtained as the output from the Mixhop GCN. The aforementioned embeddings serve as detailed representations of the spatial features that are acquired by the mixhop model for the input graphs D_{yg}.

Next, the latter component of the SGHTE model is introduced. To get precise forecasts of Bajra yield ($y_{d,t}$), it is essential to address temporal factors with the same level of consideration as spatial characteristics. Recurrent Neural Networks (RNNs) are a well-established choice for modeling sequential data due to their ability to store and update contextual information over time without relying on the restrictive Markov assumption [34]. The basic RNNs are enhanced by standard LSTMs that incorporate gated mechanisms to more effectively capture long term dependencies as well as address the problem of vanishing or exploding gradients. Yet, by their nature, they assume that time periods between observations are constant, which is hardly the case in agricultural data, where such variables as rainfall, irrigation, and fertilizer application tend to have uneven time periods. The first factor that needs to be taken into account is the fact that they should consider the temporal irregularities. In the present scenario, it is seen that the yearly rainfall data exhibits non-uniformity across a certain time frame, as detailed in Section 1.2 and shown in Figure 5. This problem is addressed by including a t-LSTM [2] to account for the potential impact of the uneven time between events. The main objective is to utilize the t-LSTM model illustrated in Figure 6 and produce the final yield $y_{d,t}$ from the embeddings $H^{(1)}$ obtained from equation (2), for the 13 years using the mixhop GCN model. The key difference in

t-LSTM compared to traditional LSTM involves modifying the memory of the prior cell state according to the elapsed time between the events.

The distinguishing change in t-LSTM is the modification of the previous cell memory $c^{(t-1)}$ state by factoring in the variable time difference. This modified part of the LSTM can be seen in Figure 6 by the blue lines of the temporal module of SGHTE. The t-LSTM first from the network calculates the short-term memory state $c_S^{(t-1)}$ which is given by,

$$c_S^{(t-1)} = \tanh\left(W_d c^{(t-1)} + b_d\right) \quad (8)$$

This short-term memory considers in the irregular time interval between the events like rainfall. Then memory discount is applied to this state, and the discounted short-term memory is given by,

$$\hat{c}_s^{(t-1)} = c_S^{(t-1)} \odot decay(\Delta^{(t)}) \quad (9)$$

The decay of this elapsed time is given by $decay(\Delta^{(t)}) = \frac{1}{\log(e + \Delta^{(t)})}$. Now finally the adjusted previous state memory $c_*^{(t-1)}$ is computed using the discounted short-term memory and the long-term memory $c_T^{(t-1)} = c^{(t-1)} - c_S^{(t-1)}$. The adjusted previous state memory is computed as,

$$c_*^{(t-1)} = c_T^{(t-1)} + \hat{c}_s^{(t-1)} \quad (10)$$

The t-LSTM only differs in this modification of the prior memory state, and the rest of the components of the t-LSTM are similar to vanilla LSTM. After this modification of the previous state, the t-LSTM follows the regular vanilla LSTM where firstly, the forget gate $f_g^{(t)}$ controls how much of the previous cell state $c^{(t-1)}$ should be retained. Then the input gate $i_g^{(t)}$ controls the amount of the $H^{(l)}$ matrix to be used as the new input to update the cell state. Candidate state $\tilde{c}^{(t)}$ refers to the possible new values that may be included in the cell state memory. Combining the prior cell state $c^{(t-1)}$ with the new candidate's memory $\tilde{c}^{(t)}$ the updated new cell state memory is $c^{(t)}$. Finally, the hidden state h^t is computed using the revised cell state. The output gate $o_g^{(t)}$ regulates the proportion of the cell state that should be sent as output. The below equations give how these gates are calculated.

$$f_g^{(t)} = \sigma(W_f H^{(l)} + U_f h^{(t-1)} + b_f) \quad \text{(forget gate) (12)}$$

$$i_g^{(t)} = \sigma(W_i H^{(l)} + U_i h^{(t-1)} + b_i) \quad \text{(input gate) (13)}$$

$$\tilde{c}^{(t)} = \tanh\left(W_c H^{(l)} + U_c h^{(t-1)} + b_c\right) \quad \text{(candidate memory gate) (14)}$$

$$c^{(t)} = f_g^{(t)} \odot c^{(t-1)} + i_g^{(t)} \odot \tilde{c}^{(t)} \quad \text{(current cell state) (15)}$$

$$o_g^{(t)} = \sigma(W_o H^{(l)} + U_o h^{(t-1)} + b_o) \quad \text{(output gate) (16)}$$

$$h^t = o^t \odot \tanh\left(c^{(t)}\right) \quad \text{(hidden state) (17)}$$

In these equations $\{W_f, W_i, W_c, W_o\}$, $\{U_f, U_i, U_c, U_o\}$, and $\{b_f, b_i, b_c, b_o\}$ are the network parameters of the forget, input, candidate, and output gates. $\{W_d, b_d\}$ are the parameters of the decomposition subspace. Using this t-LSTM model is explained in equations from (8-17).

Algorithm 2. Algorithm for spatial and temporal models of SGHTE

Input: G_T, l, P *and* Δt
Output: h^t **Predicted yield**

1. Initialize mixhop GCN with model parameters.
2. Let B = $E^{(l-1)}$) and $H^{(l)} = 0$
   For epoch 1 to 100
   For t = 1 to t
   For l = 1 to l
   For j = 1 to P
3. $B = \tilde{A}_{d,t} . B$
   If $j \in P$ **then**
4. $Q_j = B . W_j^{(l)}$
   End if
   End for
5. $E^{(l)} = \|_{j \in P} Q_j$
6. $H^{(l)} = H^{(l)} \odot E^{(t)}$
7. Calculate $f_g^{(t-1)}, i_g^{(t-1)}, \tilde{c}^{(t)}$, and $o_g^{(t)}$ using $H^{(l)}$ and equation 12,13,14, and 16
8. Modify current state memory $c^{(t)}$ using equations
9. Calculate h^t from equation 17

End for
End for

This model is trained over and the finally obtained predicted yield ($\hat{y}_{d,t}$) is,

$$\hat{y}_{d,t} = FCN(h^t) = FCN\left(o^t \odot \tanh\left(c^{(t)}\right)\right) \quad (18)$$

$\hat{y}_{d,t}$ is the predicted yield by the proposed SGHTE model. This predicted yield is now compared with the true yield for Bajra crop and the performance of the model is evaluated. In the next section design and settings for the experimental procedure of the SGHTE model is elaborated.

EXPERIMENTAL DESIGN AND SETTINGS

In this part, the authors present the experimental framework that they employed to measure the performance of the proposed model. It will contain the description of the dataset, the process in which the data collection was performed, the evaluation metrics which were used to evaluate the accuracy of the model, the training and testing process and the parameter settings. This holistic method provides the ability to replicate the results and to perform an accurate estimation of the performance of the model.

Dataset Description

The present study comprises data collected from many sources, including the DACNET website of the Ministry of Agriculture and Farmers' Welfare [38], as well as the agricultural site of Rajasthan [38]. The dataset covers 32 administrative districts (geographic resolution) of the state of Rajasthan, India, over a continuous time span from 2007 to 2020 (14 years), specifically focusing on the Kharif season crop known as Bajra (Pearl Millet). This results in a total of 416 district–year records, offering both spatial and temporal diversity in

the observations. Figure 8 shows an extract of Bajra dataset with such parameters as Area, Production, Yield, Canals, Tanks, Wells, Soil type, Phosphorus and Potassium content, Fertilizer, Hybrid seeds, Saline Soil, Alkaline Soil, Rainfall, Climate type, year, and District.

The selection of the aforementioned attributes is based on their significant contribution to the prediction of Bajra yield. As an example, the area allocated for cultivating a particular crop is denoted by the "Area." The attributes Cannals, Tanks, and Wells give details of the size of area that has been irrigated assisted by the various sources of water. The attribute Fertilizer refers to the amount of fertilizer applied expressed in tonnes. Rainfall is the average amount of meteorological precipitation taken and measured in millimetres (mm). The addition of soil nutrients including phosphorus and potassium are the main features of the soil. The salinity of soil is an important feature that can be used to refer to high levels of dissolved salts in the soil water. Considerable amounts of salt soil have been reported to have negative effects on the growth of plants which in turn results

District	Year	Area (Hectare)	Production (Tonnes)	Yield (Tonnes/Hectare)	Soil type 1	Soil type 2	Phosphorus	Pottasium	Saline Soil (Ha)	sodic or Alkaline Soil (Ha)	Annual Normal Rainfall (mm)	Climate Type	irrigated area by canals(hectare)	irrigated area by tanks(hectare)	irrigated area by wells(hectare)	hybrid seeds(quintal)	total fertilizer (tonnes)
1.AJMER	2011-12	74554	95330	1.28	lithosolsat foot	Sierozens	M	M	16712	19830	602	SEMI-ARID	4750	5256	0	2302	12649
2.ALWAR	2011-12	267792	552935	2.06	Alluvial prone t	Alluvial deposites ca	L	M	15976	97625	657	SEMI-ARID	488	0	0	10200	31421
3.BANSWARA	2011-12	101	127	1.26	Predominantly	welldrained calcare	M	M	2131	2130	950	SEMI-HUMID	59019	5005	0	0	20749
4.BARAN	2011-12	2849	3737	1.31	Clay Loam	Black of alluvial orig	M	H	1008	1584	874	HUMID	67280	5200	0	36	23249
5.BARMER	2011-12	878742	576877	0.66	Desert soils	sand dunes aeolian	M	M	1596	1989	266	ARID	1299	0	0	4430	6152
6.BHARATPUR	2011-12	118162	231978	1.96	Alluvial prone t	Alluvial deposites ca	L	M	32613	45217	664	SEMI-HUMID	3233	0	0	4580	29337
7.BHILWARA	2011-12	2882	2380	0.83	lithosolsat foot	alluvials in plains	M	H	27950	13470	683	SEMI-ARID	14624	5746	0	40	16522
8.BIKANER	2011-12	206510	200220	0.97	Desert soils	sand dunes aeolian	M	M	14134	14033	243	ARID	151568	0	0	437	9500
9.BUNDI	2011-12	2426	3182	1.31	Clay Loam	Black of alluvial orig	M	H	6009	9229	773	SEMI-ARID	113849	1203	0	137	19931
10.CHITTORGARH	2011-12	27	34	1.26	lithosolsat foot	alluvials in plains	L	M	17720	11397	842	MEDITERRANEAN	10445	936	0	0	28747
11.CHURU	2011-12	404192	293961	0.73	Desert soils	sand dunes aeolian	M	H	14134	250	355	SEMI-ARID	4265	0	0	11410	4416
12.DAUSA	2011-12	138706	228025	1.64	lithosolsat foot	Sierozens	L	H	4056	38437	561	MEDITERRANEAN	148	0	0	4112	17557
13.DHOLPUR	2011-12	82914	184950	2.23	Alluvial prone t	Alluvial deposites ca	M	L	5373	20121	745	SEMI-HUMID	8352	192	0	2686	12300
14.DUNGARPUR	2011-12	124	155	1.25	Predominantly	welldrained calcare	M	M	2819	3928	729	HUMID	9295	2489	0	0	6532
15.GANGANAGAR	2011-12	5858	10725	1.83	high soluble salts & exchangeable sodium	Alluvial deposites ca	M	H	14214	6517	274	ARID	595624	0	0	27	41416

Figure 8. Sample extract of the raw Bajra yield dataset before preprocessing, showing district-wise attributes across multiple years.

to a decline in agricultural productivity and worst, death of the plant in severe cases. The alkaline nature of the soil is a pointer of the levels of salts, which adversely affect the soil structure thereby decreasing crop production. Hybrid seeds: These are the genetically modified seeds which are utilized in the production of Bajra in quintals. The soil type and climate type are the particular composition of soil and the environment of various areas of existence.

Limitations on scope of data: The dataset is limited to Bajra crop in Rajasthan, future work will extend to multiple states and crops for broader applicability. District-level aggregation of some attributes may overlook within-district variability. Some district–year records were removed due to missing values, potentially biasing results, and label-encoding with scaling before the train–test split may cause data leakage. Rainfall parsing errors and similarity-based rather than geographic graph construction could also affect accuracy.

Data Collection and Preprocessing

The overall flow for data preprocessing is highlighted in Figure 9. The yield data is scraped from the DACNET website, which is maintained by the Ministry of Agriculture and Farmers' Welfare [38]. The provided dataset encompasses 32 districts in the state of Rajasthan over a period of 23 years. The gathered characteristics include area, production, year, and district name. The yield is computed by dividing the production data by the harvested area, as shown in columns 'Production' and 'Area'. The meteorological data, annual rainfall and the climatic type are undertaken. The dataset for annual rainfall is sourced from the official website of the Department of Water Resources, Rajasthan Government [40]. The values for the annual net irrigated area derived from water sources such as wells, canals, and tanks, as well as the quantities of hybrid seeds and fertilizers used were acquired from the Government of Rajasthan Directorate of Economics & Statistics [38]. The dataset was spread across many PDF files, whereby each file included the data corresponding to a certain year. Before analysis, each of these files underwent parsing in Python using packages such as PyPDF2 and regex. This study utilizes climatic data from reference [41] to include yield forecast for each district in Rajasthan. In accordance with [41], the geographical region of Rajasthan is classified into five distinct climatic categories, namely desert, semi-arid, semi-humid, humid, and Mediterranean. The climatic classification of each of the 32 districts was determined based on this categorization model. This has been done through regular expression functions whereby each district is assigned the respective climate. The following columns contain the categorical values: the Phosphorus, Pottasium, Soil type 1, Soil type 2, and Climate Type. The columns bearing the labels Phosphorus and Pottasium have the numerical data indicating the quantity of the elements in the soil of the sample like very low (VL), low (L), medium (M), and high (H). The columns that are marked as Soil type 1 and Soil type 2 contain the following classifications lithosols at the foothills, alluvial, prone to water logging, mostly of the reddish medium texture, clay loam, desert soils and so on. These categories are the pointers to the many different mixes of soils in each district. The climatic type includes numerous values, including dry, semi-arid, semi-humid, humid, and Mediterranean as the indicators of the climatic type of each region. Figure 10 is the sample of data that has been pre-processed.

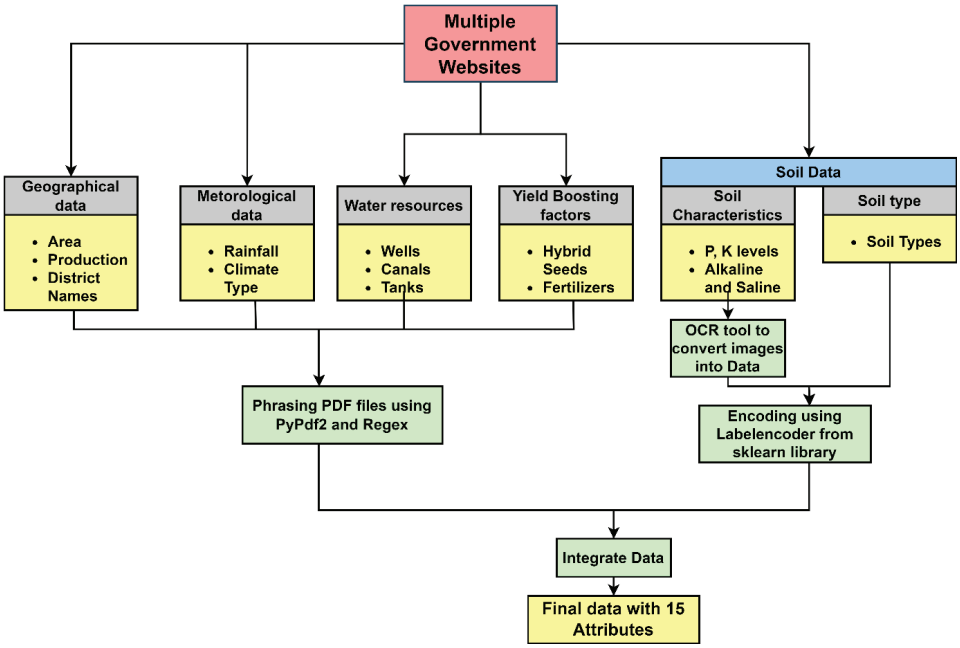

Figure 9. Flow diagram of the data preprocessing procedure, including handling of missing values, normalization, and encoding of categorical variables.

Districts	Year	Area (Hectare)	Production (Tonnes)	Yield (Tonnes/ Hectare)	Soil type 1	Soil type 2	Phosphorus	Potassium	Saline Soil Ha	Alkaline Soil
2.ALWAR	2011-12	0.264771	0.742558	0.872881	0	0	0.25	0.666667	0.000283	0.831
3.BANSWAR	2011-12	9.69E-05	0.000167	0.533898	0.428571	1	0.5	0.666667	3.74E-05	0.017926
4.BARAN	2011-12	0.002814	0.005015	0.555085	0.142857	0.142857	0.5	0	1.75E-05	0.013277
5.BARMER	2011-12	0.868836	0.774711	0.279661	0.285714	0.857143	0.5	0.666667	2.79E-05	0.016725
6.BHARATPU	2011-12	0.116827	0.31153	0.830508	0	0	0.25	0.666667	0.000578	0.384782
7.BHILWARA	2011-12	0.002847	0.003192	0.351695	1	0.571429	0.5	0	0.000495	0.114478
8.BIKANER	2011-12	0.20418	0.268881	0.411017	0.285714	0.857143	0.5	0.666667	0.00025	0.119272
9.BUNDI	2011-12	0.002396	0.004269	0.555085	0.142857	0.142857	0.5	0	0.000106	0.078369
10.CHITTOR	2011-12	2.37E-05	4.16E-05	0.533898	1	0.571429	0.25	0.666667	0.000314	0.096828
11.CHURU	2011-12	0.399634	0.39477	0.309322	0.285714	0.857143	0.5	0	0.00025	0.001919
12.DAUSA	2011-12	0.13714	0.306221	0.694915	1	0.428571	0.25	0	7.16E-05	0.327055
13.DHOLPUF	2011-12	0.081977	0.248374	0.944915	0	0	0.5	0.333333	9.49E-05	0.171107
14.DUNGARI	2011-12	0.00012	0.000204	0.529661	0.428571	1	0.5	0.666667	4.96E-05	0.033235
15.GANGAN	2011-12	0.005789	0.014399	0.775424	0.857143	0	0.5	0	0.000252	0.055278
16.HANUMA	2011-12	0.072056	0.135357	0.584746	0.857143	0	0.5	0.666667	0.001316	1
17.JAIPUR	2011-12	0.313714	0.663726	0.661017	1	0.428571	0.5	0.666667	0.001316	0.003707
18.JAISALME	2011-12	0.141881	0.098669	0.216102	0.285714	0.857143	0.5	0.666667	0.000221	0.1998

Alkaline Soil	Rainfall	Climate type	Cannals(Hectare)	Tanks(Hectare)	Wells(Hectare)	Hybrid seeds(Quintal)	Fertilizer(Tonnes)
0.831	0.075136	0.75	0.00076	0	0.885893	0.711744	0.423656
0.017926	0.108687	1	0.09186	0.174817	0.02685	0	0.267276
0.013277	0.099984	0.25	0.104717	0.181628	0.434257	0.002512	0.30391
0.016725	0.030365	0	0.002022	0	0.335507	0.30912	0.053382
0.384782	0.075938	1	0.005032	0	0.647739	0.319587	0.393119
0.114478	0.078114	0.75	0.022761	0.200699	0.309862	0.002791	0.205337
0.119272	0.027731	0	0.235906	0	0.273553	0.030493	0.102441
0.078369	0.088419	0.75	0.177199	0.042019	0.200551	0.00956	0.25529
0.096828	0.09632	0.5	0.016257	0.032693	0.35989	0	0.384473
0.001919	0.040556	0.75	0.006638	0	0.174166	0.796176	0.027944
0.327055	0.064144	0.5	0.00023	0	0.309451	0.28693	0.220503
0.171107	0.085213	1	0.012999	0.006706	0.206195	0.187426	0.14347
0.033235	0.083381	0.25	0.014467	0.086937	0.053274	0	0.05895
0.055278	0.031281	0	0.927053	0	0.001821	0.001884	0.570116
1	0.064487	0	0.593005	0	0.003272	0.179541	0.427422
0.003707	0.021204	0.75	0.01285	0.020293	0.58722	0.705882	0.387111
0.1998	0.042273	0	0.116294	0	0.087365	0.092108	0.046627

Figure 10. Example of encoded dataset after preprocessing, ready for input into the SGHTE model.

Data pre-processing like removing the rows with null values and integer-encoding the categorical values in order to transform the input data to a suitable format is necessary for machine learning models. The pre-processing steps involved are:

- Pandas' library is used to import the data shown in Figure 10.
- PyPDF2 is utilized to phrase rainfall, irrigated sources area, hybrid seeds, and fertilizer data
- The soil data was present in images. The data was extracted from the images using Microsoft's OCR tool, and stored in CSV files, which were added to the current dataset.
- Encoding of categorical values is performed using the LabelEncoder package from sklearn library.
- Scaling of the data was performed with StandardScaler and MinMaxScaler package from sklearn library
- Finally, the null values are removed with the help of Pandas library

After preprocessing, the yield data from 2010-2020 for 32 districts is considered for graph construction as shown in Figure 6. The yield prediction model proposed in Figure 6, is trained over this data and the settings for this are discussed in the next subsection.

Training and Testing Procedure With Hyperparameter Settings

The hyperparameter settings of the Mixhop GCN model and t-LSTM are discussed in this section. The

parameter setting for generating the year-wise embeddings ($H^{(1)}$) as shown in Figure 6 are: number of features are 15 (which represents the number of input features (x) for each district in the graph after preprocessing explained in above Section), 32 Hidden Channels, dropout Probability is 0.5, the learning rate for the Adam optimizer is 0.001. Based on the Figure 6, the t-LSTM model is trained on $H^{(1)}$ with the window size of 3, and the input, hidden, output sizes are 32,64, and 32, respectively. The training of both mixhop GCN and t-LSTM is conducted prior to testing. Initially, the Mixhop model undergoes training, which involves:

- Initializing the Mixhop GCN model with three GCN layers (conv1, conv2, conv3), where each layer will aggregate information from 0-hop (self), 1-hop, 2-hop, and 3-hop neighbours. A dropout layer for regularization is also included. This generates the year-wise embeddings as shown in Figure 6.
- The training process iterates over 100 epochs. The model processes the data of every year (spatial data of districts) one at a time in each epoch.
- A mixhop GCN model is applied to the data of each year, which will produce node embeddings (spatial features) of the districts. These embeddings are gathered and also to determine the RMSE loss between the embeddings and the actual yield values.
- The annual loss is calculated with the help of RMSE and the mean loss on the years is calculated. This is followed by the loss being backpropagated to adjust the model parameters using the optimizer.

The steps involved in the training of the t-LSTM model are:

- The t-LSTM model takes the embeddings created by mixhop GCN as its input and the time series data is divided into training and testing sets. This study uses a 70:30 data division i.e. 70 percent of the data is taken as a training data and 30 percent is taken as a testing data.
- The t-LSTM model is trained with a fully connected layer with learning rate 0.001. The model is defined by processing the data in 3 sliding window, with each window being a sequence of embeddings across a series of time steps.
- For each test year, the TLSTM model will use input sequences generated from prior years. After training is completed the SGHTE model is evaluated on the test data. Using the same window sliding approach the model is iterated on the testing data. Finally, evaluation metrics are calculated.

Details on Evaluation Parameters

Three evaluation measures, namely *RMSE*, R^2, and Pearson correlation coefficients, are used to assess the performance of the SGHTE model. The quantifies the mean of the deviations between the projected and actual yield. The metric provides an indication of the degree to which the model's predictions align with the observed data, with lower values indicating superior performance. Equation (6) can be used to calculate RMSE. coefficient of determination is an important parameter in regression analysis, which is used to estimate the level of goodness of fit to a model. R^2 is defined to be the portion of the variation in the dependent variable which can be predicted by the independent variables. It has a value between 0 and 1 with a larger value having a better fit. Equation (13) can be used to estimate R^2. Here $y_{d,t}$ is an average of the actual yield.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_{d,t} - \hat{y}_{d,t})^2}{\sum_{i=1}^{m}(y_{d,t} - \bar{y}_{d,t})^2} \tag{13}$$

The Pearson correlation coefficient is used to determine the direction and strength of the linear relationship between two variables. It has the value of -1 to 1, a value of one to -1 indicates how strong the linear relationship is, and a value of zero to -1 indicates how weak the linear relationship is. The value of positive correlation is also high and this implies that the model has a high adherence to the actual yield values implying good model performance. According to equation (14), correlation coefficient (r) is provided. A correlation analysis can be used to determine which features (rainfall, temperature, soil type, etc.) are most related to crop yield to help select features to include in the model $y_{d,t}$ is the average of predicted yields.

$$r = \frac{\sum_{i=1}^{m}(y_{d,t} - \bar{y})(\hat{y}_{d,t} - \bar{\hat{y}}_{d,t})}{\sqrt{\sum_{i=1}^{m}(y_{d,t} - \bar{y}_{d,t})^2 \sum_{i=1}^{m}(\hat{y}_{d,t} - \bar{\hat{y}}_{d,t})^2}} \tag{14}$$

Table 3 summarizes the experimental setup of the proposed SGHTE model, which makes the methodology completely transparent as to be reproducible. The algorithmic process included these stages: data cleansing, categorical encoding, scaling, and consolidation of heterogeneous data into a single data set of 15 attributes of 32 districts in 13 years. The spatial input of the Mixhop GCN module was constructed by creating the graphs every year with cosine similarity between the district-level feature vectors. This spatial module used three convolution layers with multi-hop neighbourhood aggregation and then a t-LSTM time based module that was used to forecast irregularities in time dependence in attributes like rainfall. Each module was trained in 100 epochs with a sliding window training method with a 70:30 split between training and testing. RMSE was used as the primary loss function, while RMSE, R^2, and Pearson's correlation coefficient served as evaluation metrics. These details collectively provide a comprehensive blueprint for reproducing the results.

After setting up the model the experimentation is performed and predicted yield results are obtained. These results are discussed in the following section.

RESULTS AND DISCUSSION

The results of the experimentation for the proposed SGHTE model are divided into three parts: 1) Sensitivity analysis, where the optimal hyperparameter setting of the

Table 3. Model architecture, hyperparameters, preprocessing steps, and training/testing settings for the SGHTE model

Aspect	Details
Data Preprocessing	Removal of rows with null values (Pandas), integer encoding of categorical variables (LabelEncoder, sklearn), scaling using StandardScaler and MinMaxScaler (sklearn), OCR extraction for soil data (Microsoft OCR), PDF parsing for rainfall, irrigation, hybrid seeds, and fertilizer data (PyPDF2, regex).
Dataset	DAWNET-based data were used as the source of 20072012-2017-2020 Bajra crop statistics of 32 districts of Rajasthan sourced through DACNET, Department of water resources, Rajasthan, Directorate of Economics and statistics and literature-based climate classification. Finally data set: 15 variables, 416 entries, time-varying (temporal) and spatial (spatial) data.
Graph Construction	Nodes: 32 districts; Edges: cosine similarity-based weights between district feature vectors; Graphs constructed yearly from adjacency matrices A^t.
Model Architecture	**Spatial module:** Mixhop GCN with 3 layers (conv1, conv2, conv3), aggregating 0-hop, 1-hop, 2-hop, and 3-hop neighbours; dropout layer (0.5). **Temporal module:** t-LSTM with input size 32, hidden size 64, output size 32, incorporating time-aware memory decay. Fully Connected Layer (FCN) for final yield prediction.
Hyperparameters – Mixhop GCN	Input features: 15; Hidden channels: 32; Dropout: 0.5; Learning rate: 0.001 (Adam optimizer); Epochs: 100.
Hyperparameters – t-LSTM	Window size: 3; Input size: 32; Hidden size: 64; Output size: 32; Learning rate: 0.001; Epochs: 100.
Training/Testing Split	70% training, 30% testing (sliding window approach).
Loss Function	RMSE (Root Mean Square Error).
Evaluation Metrics	RMSE, R^2 (Coefficient of Determination), Pearson's correlation coefficient (r).

model is selected. 2) Predicted results, where the predicted output of the model is compared to true yield, and 3) State-of-the-art comparison, where the proposed model and its variants are compared with other effective crop yield models.

Sensitivity Analysis

The proposed model's performance is sensitive to mixhop GCN parameter and learning rate =[0.0001,0.001,0.01,0.1] The optimal tuning of these sensitive parameters is required for optimal performance metrics.

In the map given for illustration purposes in Figure 3, for Jaipur, beyond three hops , the relevance of neighboring districts typically diminishes. The districts that would fall into fourth-order neighbours are often too far geographically and may not have significant socio-economic or cultural ties to the reference district. Thus, here testing up to three hops is evaluated, focusing on districts where influences are most likely to be felt. The learning rate is also evaluated for varying range from 0.0001 to 0.1.

Figure 11 presents a comparative bar plot illustrating the performance metrics of the SGHTE model by varying the MixHop GCN hyperparameter which is a set of integer adjacency powers that refer to the hops over which the model aggregates information. The learning rate is kept fixed at a value of 0.001. A value of equal to 0 implies that the model simply combines information from its immediate neighbors, which is analogous to a conventional GCN. From Figure 11, it is evident that as the hyperparameter set includes more hops (e.g., [0,1,2,3]), the performance improves. The RMSE values remain low, while the R^2 and correlation values are relatively high, especially in the hop set [0,1,2,3], which seems to give the best overall results with an R^2 of 0.6775 and a of 0.8912. This implies that adding a larger range of hops, i.e. considering districts linked by distant districts rather than simply their immediate neighbours, leads to more accurate and trustworthy predictions by the Mixhop GCN module in SGHTE model.

Figure 12 consists of three column graphs that depict the model assessing performance metrics: Test shown in blue, Test shown in red, and the Test is shown in green. The x-axis indicates various learning rates. The y-axis corresponds to the numerical values of the performance metrics. The optimal hop value from the above experiment in Figure 12 is used for learning rate sensitivity analysis.

The best learning rate of 0.001 would give the minimum RMSE of 0.1306, and it means that the model would be able to train without overfitting or underfitting data sets. The reduced values of RMSE show improved model accuracy. R 2 value of 0.6775, which is the largest at the learning rate of 0.001, implies that this learning rate allows the model to best represent the variation in the Bajra yield data. The highest correlation coefficient of 0.8912 is achieved when learning rate is 0.001 indicating that learning rate provides the best compromise of accurate predictions with strong correlation. The larger values of the correlation coefficient and R 2 denote the superiority of the model. This indicates that with a learning rate of 0.001, the model is successfully extracting knowledge of the relational patterns in the

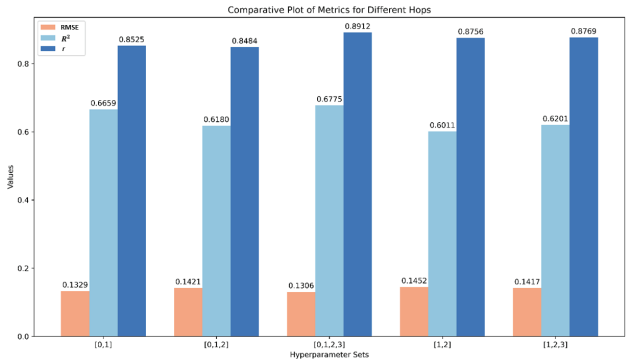

Figure 11. Effect of varying the Mix-hop hyperparameter on model performance.

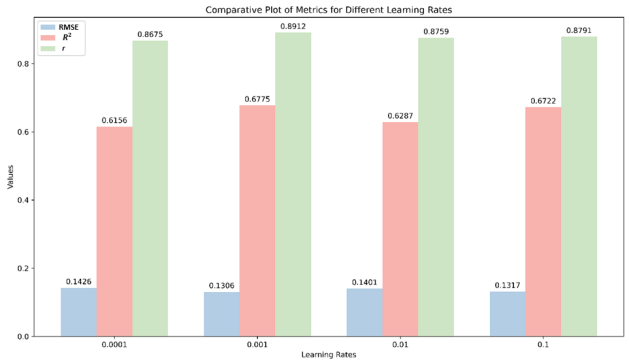

Figure 12. Impact of different learning rates on proposed model.

district-wise data without being too aggressive in weight adjustments (which may result in overfitting) or being too overly conservative in weight adjustments (which may result in underfitting).

Comparative Analysis

This section evaluates the performance of the proposed model by analysing true versus predicted yield, conducting a comparative analysis with state-of-the-art models, and examining model variations through feature reduction studies.

Figure 13 shows the true yield values vs the predicted yield values of all the 32 districts for the final year of data 2019-2020. The model was trained on the yields for all the previous years and the targeted year predicted yields for the year 2019-2020 are given in the Figure 14. The model has a high predictive level of the yields in all 32 districts, and the margin of error is rather low. In training, the RMSE is decreased to 0.1306 which is a better predictive result. The blue dashed line represents the real yield values of the districts and the red solid line is the projections of the model. The fact that in the majority of the districts the RMSE of the actual and predicted lines is smaller proves that the model is a good reflection of the spatial and temporal patterns. This illustrates the ability of SGHTE model to represent Mix-hop spatial interactions and temporal patterns that result to accurate yield predictions.

The proposed model is compared with its variants alongside with other state-of-the-art models to show the improvement in yield prediction. The modified versions of the proposed model for component wise analysis are:

- The proposed model (SGHTE) is trained without the features and denoted as LASGHTE (Limited Attribute SGHTE)
- In the proposed model seen in Figure 6, the irregularity aware temporal part is replaced with traditional LSTM (MGCN+LSTM).
- Mutli-hop spatial part is replaced with traditional GCN (GCN+t-LSTM).
- Irregularity aware temporal part is replaced with GRU (MGCN+GRU).
- Both spatial and temporal parts are replaced with traditional GCN and LSTM (GCN+LSTM).

All comparisons are performed after optimizing the model parameters, ensuring consistency by utilizing the same datasets.

A comparison of the predicted vs. actual yield values for the proposed SGHTE model and component wise analysis is provided in Figure 14. This figure illustrates the regression analysis of predicted vs. actual yield values along with their regression coefficients (R values). The red line represents the regression line, showing the model's overall trend in predicting yield values. It is a line of best fit for the predicted yield (y-axis) against the true yield (x-axis) in Figure 14. As compared to the shaded red region around the regression line, which denotes the uncertainty or variability of the predictions of the model. When the spread is narrow it means that there is more confidence in the predictions whereas when the spread is wide the prediction is less certain. The blue dots, that are presented in Figure 14, point to each single data point, and they plot the actual yield (x-axis) against the yield predicted by the model (y-axis).

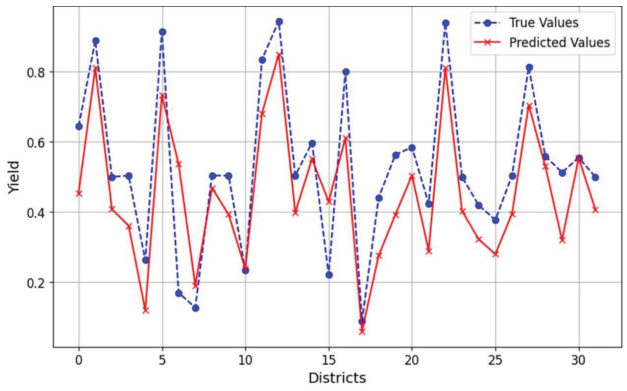

Figure 13. Comparative plot of actual versus predicted Bajra yield using the proposed model, showing high alignment and predictive accuracy.

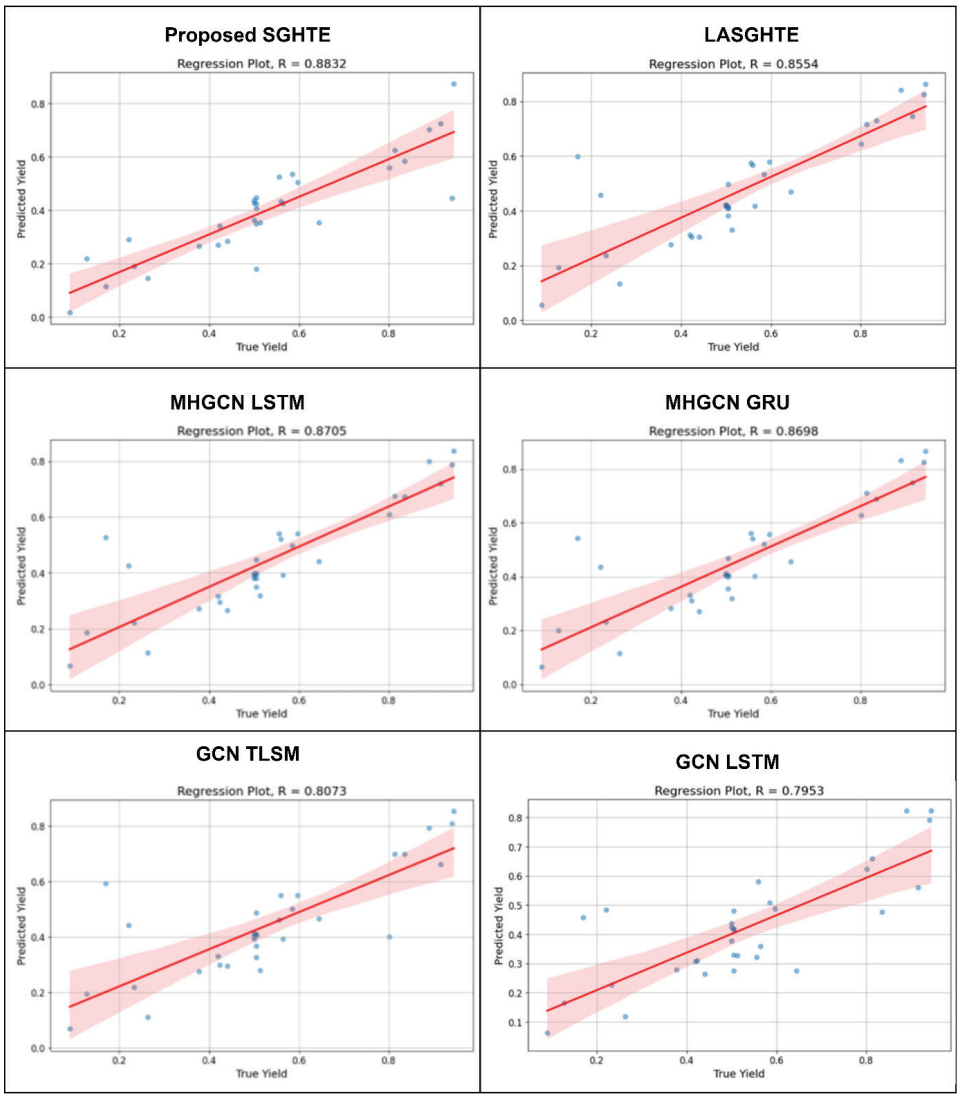

Figure 14. Regression plot of the proposed SGHTE model along with component-wise analysis, demonstrating the contributions of spatial and temporal modules to prediction performance.

The closeness of the blue dots to the red line is the extent to which the model predictions match the actual ones.

The proposed model shows highest regression coefficient $R = 0.8832$, indicating strong predictive accuracy for crop yield. The red regression line closely corresponds with the actual yield values, but the blue dots are densely grouped around it, indicating minimal variation and enhanced yield prediction. This is because of the depth wise district connections based on regional dynamics and handling temporal irregularities with the help of t-LSTM. However, when LASGHTE is evaluated, it shows a slightly lower $R = 0.8554$ compared to the proposed SGHTE, indicating a small decrease in predictive accuracy. This variant excludes certain features (like yield boosters and water resources), which leads to a reduction in the predictive power. This indicates that the inclusion of fertilizers and seed information helped with the yield prediction improvement of 3.2% in the proposed model compared to LASGHTE.

The variant MHGCN LSTM shows R-value of 0.8705, close to that of the proposed SGHTE, indicating a strong predictive alignment, though slightly less than the proposed model. Here, the t-LSTM component is replaced with a traditional LSTM and MixHop is retained, indicating that model still captures complex spatial dependencies well keeping the R high, but do not account for irregular temporal variations as explained in Figure 7, this is why it shows 0.19% decrease in R value compared to proposed model. Extension of LSTM model, GRU is also applied for ablation study, it shows R-value of 0.8698, similar to MHGCN LSTM, indicating a strong predictive performance. Replacing t-LSTM with GRU still yields good results, but GRU might not handle irregular temporal data as effectively as t-LSTM. While GRU is generally capable in time-series modelling, it

may introduce a small bias, showing 0.015% of decrease in R compared to proposed model.

GCN t-LSTM shows lower regression coefficient of R = 0.8073, indicating a weaker alignment between predicted and actual yields. By replacing MixHop GCN with a traditional GCN, this model likely captures fewer complex spatial relationships, which reduces its predictive capability. However, the t-LSTM component still helps in managing irregular temporal patterns, which keeps the regression coefficient from dropping further. Whereas the variant GCN-LSTM shows R = 0.7953 suggesting weaker alignment, with more bias and error in predictions.

State-of-art models comparison (SOTA)

The performance of proposed model is compared with the state-of-the-art models used for yield prediction like LSTM [20], CNN-RNN [21], and GNN-RNN [32]. Commonly used for time-series forecasting models in crop yield prediction is LSTM [20], which captures long-term temporal relationships. CNN-RNN [21] and GNN-RNN [32] models are the ensemble models that capture both spatial and temporal features and these architectures provide a strong comparison point for the SGHTE model with GNN-RNN [32] model showing improvements in using graph based spatial feature extraction [43]. These state-of-the-art models are evaluated using the pre-processed Bajra crop data, maintaining all the parameters and methodology approaches described in the study identically.

Table 4 provides a comparison of the performance of various state-of-the-art models, including the proposed MixHop GCN with Time-aware LSTM (denoted as MGCN t-LSTM), across three metrics: ,, and (pearson correlation). The proposed model outperforms the recent models with the best ,, and CORR scores of 0.1306, 0.6775, and 0.8912, respectively. Table 4 shows that the SGHTE achieves high performance compared to the LASGHTE variation as it includes irrigation accessibility, fertilizer, hybrid seeds, and soil types, that impacts significantly on the crop yield. The model where LSTM is used for temporal variations MGCN LSTM and LSTM [20] shows high RMSE, which is 11.83% and 23.7% compared to proposed model. Whereas MGCN LSTM and LSTM [20] shows low R^2 and r which is 1.23%,1.64% and 5.54%, 1.26% compared to proposed model respectively. This is because the time irregularity employed in proposed model helps capture temporal patterns more effectively than the standard LSTM.

Table 4. Comparison of proposed model with SOTA models

Model	Improved Attribute Dataset		
	RMSE	R^2	*r*
SGHTE (Proposed)	**0.1306**	**0.6775**	**0.8912**
LASGHTE	0.1549	0.5459	0.8662
MGCN LSTM	0.1482	0.6692	0.8766
MGCN GRU	0.1546	0.6062	0.8686
GCN t-LSTM	0.1420	0.6187	0.8075
GCN LSTM	0.1492	0.5790	0.7841
LSTM [4]	0.1712	0.640	0.880
CNN-RNN [8]	0.21	0.59	0.50
GNN-RNN [32]	0.4914	0.6236	0.8549
ABP-XGBoost [44]	0.3224	0.5223	0.6905
Random Forest [45]	0.582	0.4995	0.6244

The SGHTE model incorporates a high-order spatial representation, which allows it to model crop yield predictions across districts that are not only geographically proximate but also considers shared agricultural characteristics, such as soil types, irrigation facilities, and fertilizer usage. This comprehensive spatial modelling helps SGHTE capture variations in yield due to regional factors, providing a robust context for districts that might share similar environmental and resource conditions even if they are not immediate neighbours.

The proposed SGHTE model shows the lowest RMSE, indicating superior accuracy in yield prediction. The model GNN-RNN [32] shows high RMSE value compared to the proposed model, reflecting its limitations in capturing spatial relations that extend beyond immediate geographical proximities. This leads to low shows R^2 and r which is 7.96% and 4.07% compared to proposed model. Similarly, CNN-RNN model which captures the spatial relations directly from the crop data instead of incorporating the graph structure shows high RMSE leading to 12.92% and 43.90% of decrease in and compared to proposed model.

The proposed model shows 14.54% and 12% of improvement in and compared to GCN-LSTM which captures the immediate neighbours ignoring the second-order and third order neighbours leading to inferior yield prediction. When the proposed model is compared with MGCN GRU where instead of t-LSTM, GRU is employed, it shows 10.52% and 2.54% decrease in R^2 and r. This shows that the proposed model is not only able to capture the spatial relations of district with its neighbour districts at multiple hop-distances (P = [0,1,2,3]) but also account for temporal irregularities through decay time which is not addressed through GRU. Thus, the SGHTE model achieves robust and reliable yield predictions by effectively combining high-order spatially-informed graph structures with advanced temporal processing mechanisms, significantly outperforming SOTA models.

The suggested SGHTE framework is superior to previous models that have been mentioned in the literature. An example is that older machine learning methods (Random Forest and Gradient Boosting) [12, 20] could be highly accurate in some cases, but fail to represent higher-order spatial patterns and non-periodic time scales, which tend to overfit local region-specific patterns. Deep learning methods such as ANN [21,43] and hybrid CNN–RNN architectures [5, 19, 44] addressed spatial or temporal dependencies individually, but CNNs were constrained by regular grid

structures and RNN-based models struggled with long-term dependencies or irregular event timing. Similarly, GCN-based methods [45,46] effectively modelled spatial relationships but were restricted to first-order neighbours, thereby missing broader regional influences. Conversely, Mix-hop GCN incorporated in SGHTE allows capturing multi-hop spatial dependencies when a single layer is used, whereas the t-LSTM extension has been demonstrated to be well motivated to handle non-uniform time intervals as demonstrated by an RMSE of 0.1306, a R2 of 0.6775, and a Pearson correlation of 0.8912 on the Rajasthan dataset. Compared to CNN to RNN hybrids [14,16] and GCN to LSTM models [46], SGHTE lowers the RMSE by the magnitude of 23.7% and it can obtain higher correlation values, which confirm the increased ability of this model to reflect the complex patterns of space-temporal variations. These enhancements confirm that SGHTE is not only filling the gaps that were found in the previous research but also offers a better generalizable and solid method of crop yield prediction.

CONCLUSION

The Spatial Graph Hop with Temporal Enhancement (SGHTE) model integrates Mix Hop Graph Convolutional Networks (Mix Hop GCN) with Time-Aware Long Short-Term Memory (t-LSTM) to enhance district-level yield predictions for the Bajra crop across 32 districts in Rajasthan, India. Leveraging a comprehensive dataset span from 2010 to 2020, the model incorporates key agronomic features such as soil classification, hybrid seed quantity, and irrigation sources, enhancing predictive accuracy. We propose SGHTE, a hybrid model that leverages multi-region spatial dependencies via an extended MixHop GCN and models irregular temporal patterns with a time-aware LSTM. By capturing cross-district interactions and long-range spatial–temporal correlations, SGHTE significantly improves crop yield prediction accuracy across Rajasthan's 32 districts. Spatial Graph Hop with Temporal Enhancement (SGHTE) is a model that is based on Mix Hop Graph Convolutional Network (Mix Hop GCN) and the Time-Aware Long Short-Term Memory (t-LSTM) on top of the former to predict on a district-level yield of the Bajra crop in 32 district of Rajasthan, India. The model uses several important agronomic characteristics, including soil types, quantity of hybrid seeds and source of irrigation which were obtained through a period of time of 10 years (2010-2020) which increases its predictive power. SGHTE model has been shown to be better at forecasting than the state-of-the-art models with a RMSE of 0.1306, coefficient of determination(R2) of 0.6775 and Pearson correlation coefficient of 0.8912. SGHTE outperforms a vanilla LSTM variant by an 11.84% decrease in RMSE and 1.24% increase in R 2, which reflects its efficiency in describing spatial interaction, dealing with irregularities, and including various agricultural variables, which are important in predicting crop yields in geographically and agriculturally diverse areas. The comparative analysis proves its capability of covering the complexity of agricultural data and it is therefore quite useful in data driven decision making in agriculture.

Limitations and Future Work: Although SGHTE has high performance on predicting yields of Bajra in the state of Rajasthan, its extrapolation to other crops or regions is not tested and will have to be re-trained with region-specific data. The model can be affected by absent or noisy inputs (e.g., rainfall, fertilizer use), and the MixHop GCN and t-LSTM components will cost more computationally at scale. The existing constraints are that the aggregation is at the district level, removals of missing values, and that preprocessing might lead to data leakage, and the construction of a graph is not based on strict geography, but similarity. The work will be extended into a variety of states and crops, tested in out-of-distribution regimes, use higher-resolution data, and run case studies in which feature-importance analysis is used to enhance robustness, interpretability, and policy actions. In addition to Bajra forecasting, the spatial-temporal framework of SGHTE provides flexible addressing of different problems in engineering and natural sciences.

REFERENCES

[1] Pathare AA, Sethi D. Review and Study of Smart Net Meter Unit for Solar Roof Top Photo-Voltaic (SRTPV) Systems Enabled with Internet of Things (IoT) for EV Infrastructure. Modern Computing Technologies for EV Efficiency and Sustainable Energy Integration 2025;317–340.

[2] Pathare AA, Sethi D. Development of IoT-enabled solutions for renewable energy generation and net-metering control for efficient smart home. Discov Internet Things 2024;4:11. [CrossRef]

[3] Pathare AA, Singh RP, Sethi D. An IoT-Enabled Smart Net-Metering System for Real-Time Analysis of Renewable Energy Generation in MATLAB/Simulink. J Ins Eng India Ser B 2024;105:1583–1598. [CrossRef]

[4] Forastieri V. The ILO Programme on Safety and Health in Agriculture: The challenge for the new century–providing occupational safety and health services to workers in agriculture1. Health Safety Agric 2000;1:1.

[5] Yunker JA. Economic growth in China and India: The potential role of population. World Dev Sustain 2024;4:100130. [CrossRef]

[6] Malhi GS, Kaur M, Kaushik P. Impact of climate change on agriculture and its mitigation strategies: A review. Sustain 2021;13:1318. [CrossRef]

[7] United Nations. World Population Prospects 2022: Summary of Results. Available at: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf. Acessed on 4 Apr 2026.

[8] Sisman Z, Tekiner-Mogulkoc H. Using malmquist TFP index for evaluating agricultural productivity: Agriculture of Türkiye NUTS2 regions. Sigma J Eng Nat Sci 2022;40:513–528. [CrossRef]
[9] Ansarifar J, Wang L, Archontoulis SV. An interaction regression model for crop yield prediction. Sci Rep 2021;11:1–14. [CrossRef]
[10] Satyavathi CT, Sankar SM, Singh SP, Kapoor C, Soumya SL, Ambawat S, et al. Breeding Climate Resilient Pearl Millet Cultivars for India. Springer 2024;31–55. [CrossRef]
[11] Pearl Millet (Bajra) Millet Statistics | Millet Statistics by ICAR (Indian Council of Agricultural Research) & IIMR (Indian Institute of Millet Research). (n.d.). https://www.milletstats.com/pearl-millet-bajra/ Accessed on Apr 09, 2026.
[12] Sharma SK, Sharma DP, Gaur K. crop yield predictions and recommendations using random forest regression in 3a agroclimatic zone, rajasthan. J Data Acquisition Proc 2023;38:1635.
[13] Choudhary NK, Chukkapalli SSL, Mittal S, Gupta M, Abdelsalam M, Joshi A. Yieldpredict: A crop yield prediction framework for smart farms. IEEE Xplore 2020;2340–2349. [CrossRef]
[14] Khaki S, Wang L, Archontoulis SV. A CNN-RNN framework for crop yield prediction. Front Plant Sci 2020;10:1750. [CrossRef]
[15] Qiao M, He X, Cheng X, Li P, Zhao Q, Zhao C, et al. KSTAGE: A knowledge-guided spatial-temporal attention graph learning network for crop yield prediction. Inf Sci 2023;619:19–37. [CrossRef]
[16] Sun J, Di L, Sun Z, Shen Y, Lai Z. County-level soybean yield prediction using deep CNN-LSTM model. Sensors 2019;19:4363. [CrossRef]
[17] Dimensions AI: The Most Advanced Scientific Research Database. Available at: https://www.dimensions.ai/. Accessed on 4 Apr 2026.
[18] Jhajharia K, Mathur P, Jain S, Nijhawan S. Crop yield prediction using machine learning and deep learning techniques. Proc Comput Sci 2023;218:406–417. [CrossRef]
[19] Morales A, Villalobos FJ. Using machine learning for crop yield prediction in the past or the future. Front Plant Sci 2023;14:1128388. [CrossRef]
[20] Jhajharia K, Mathur P. Machine Learning Based Crop Yield Prediction Model in Rajasthan Region of India. Iraqi J Sci 2024. [CrossRef]
[21] Vashisth A, Goyal A. Prediction of mustard yield using different machine learning techniques: a case study of Rajasthan, India. Int J Biometeorol 2023;67:539–551. [CrossRef]
[22] Sharma SK, Sharma DP, Gaur K. Machine Learning Techniques for Crop Yield Forecasting in Semi-Arid (3A) Zone, Rajasthan (India). Curr Agric Res J 2023;11. [CrossRef]
[23] Wang J, Wang P, Tian H, Tansey K, Liu J, Quan W. A deep learning framework combining CNN and GRU for improving wheat yield estimates using time series remotely sensed multi-variables. Comput Electron Agric 2023;206:107705. [CrossRef]
[24] Radhika A, Masood MS. Crop Yield Prediction by Integrating Et-DP Dimensionality Reduction and ABP-XGBOOST Technique. J Internet Serv Inf Secur 2022;12:177–196. [CrossRef]
[25] Asamoah E, Heuvelink GB, Chairi I, Bindraban PS, Logah V. Random Forest machine learning for maize yield and agronomic efficiency prediction in Ghana. Heliyon 2024;10. [CrossRef]
[26] Wang F, Li J, Peng D, Yi Q, Zhang X, Zheng J, et al. Estimating Soybean Yields using Causal Inference and Deep Learning Approaches with Satellite Remote Sensing Data. IEEE Xplore 2024. [CrossRef]
[27] Wang H, Zhang L, Zhao J. Application of a Fusion Attention Mechanism-Based Model Combining Bidirectional Gated Recurrent Units and Recurrent Neural Networks in Soil Nutrient Content Estimation. Agronomy 2023;13:2724. [CrossRef]
[28] Khaki S, Wang L. Crop yield prediction using deep neural networks. Front Plant Sci 2019;10:621. [CrossRef]
[29] Lamba V, Hooda S, Ahuja R, Kaur A. Wheat yield prediction using feedforward neural networks. IEEE Xplore 2021;1–6. [CrossRef]
[30] Dharmaraja S, Jain V, Anjoy P, Chandra H. Empirical analysis for crop yield forecasting in India. Agri Res 2020;9:132–138. [CrossRef]
[31] Fan J, Bai J, Li Z, Ortiz-Bobea A, Gomes CP. A GNN-RNN approach for harnessing geospatial and temporal information: application to crop yield prediction. Proc AAAI Conf Artif 2022;36:11873–11881. [CrossRef]
[32] Balasubramanian A, Elangeswaran SVJ. A novel power aware smart agriculture management system based on rnn-lstm. Electr Eng 2025;107:2347–2368. [CrossRef]
[33] Baytas IM, Xiao C, Zhang X, Wang F, Jain AK, Zhou J. Patient subtyping via time-aware LSTM networks. 23rd ACM SIGKDD international conference on knowledge discovery and data mining. Halifax, Canada; 2017. p. 65–74. [CrossRef]
[34] Nathawat R, Singh SK, Sajan B, Pareek M, Kanga S, Đurin B, et al. Integrating Cloud-Based Geospatial Analysis for Understanding Spatio-Temporal Drought Dynamics and Microclimate Variability in Rajasthan: Implications for Urban Development Planning. J Indian Soc Remote Sens 2025. [CrossRef]
[35] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv 2016.
[36] 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

[37] Government of Rajasthan Directorate of Economics & Statistics. Rajasthan Agriculture Statistics Agriculture Statistics-Government of Rajasthan. Available at: https://rajas.rajasthan.gov.in/Index.aspx Accessed on Apr 09, 2026.
[38] Rajasthan agriculture department https://agriculture.rajasthan.gov.in/home Accessed on Apr 09, 2026.
[39] 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 Accessed on Apr 09, 2026.
[40] Gunawat A, Dubey SK, Sharma D. Development of indices for aridity and temperature changes pattern through GIS mapping for Rajasthan, India. Clim Chang Environ Sustain 2016;4:178–189. [CrossRef]
[41] Abu-El-Haija S, Perozzi B, Kapoor A, Alipourfard N, Lerman K, Harutyunyan H, et al. Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. ArXiv 2019;21–29.
[42] Mishra A, Rawat S, Gautam S, Mishra EP. Comparison between Different Mustard Yield Prediction Models Developed using Various Techniques for Udaipur Region of Rajasthan. Int J Environ Clim Chang 2022;12:475–485. [CrossRef]
[43] Cedric LS, Adoni WYH, Aworka R, Zoueu JT, Mutombo FK, Krichen M, et al. Crops yield prediction based on machine learning models: Case of West African countries. Smart Agric Technol, 2022;2:100049. [CrossRef]
[44] Meena RP, Dhama A. Relevance of thermal units in deciding sowing time and yield prediction of groundnut (Arachis hypogaea L.) under irrigated condition of western Rajasthan. J Agrometeorol 2004;6:62–69. [CrossRef]
[45] Sharma V, Singh PK. Performance of AquaCrop model for predicting yield and biomass of okra (Abelmoschus esculentus) crop. Indian J Agric Sci 2023;93:899–905. [CrossRef]
[46] Vats S, Shankar VG, Mohapatra S, Singh H. Enhancing Crop yield through recommendation system using Fuzzy Inference Model. IEEE Xplore 2023;1–5. [CrossRef]

Symbol	Meaning	Symbol	Meaning
G_T	The yearly district attribute relational	$\|\|$	Column-wise concatenation operator
$H^{(l)}$	Output of Mixhop GCN	t	Time
k	Number of nearby nodes	Δt	Elapsed time
x	Feature/attribute matrix	FCN	Fully connected layer
A_t	Adjacency matrix each year	V	Graph nodes
D	Degree matrix of the adjacency matrix	$\mathcal{E}$	Node edges
I	Identity matrix	N	Number of districts
$\tilde{A}_t$	Normalized adjacency matrix with self-connections	$c_S^{(t-1)}$	Short-term memory gate
W_n^j	Trainable weight matrix	$\hat{c}_s^{(t-1)}$	Discounted short-term gate
l	Number of layers of mixhop GCN	$c_T^{(t-1)}$	Long-term memory gate
$y_{d,t}$	Actual yield	$c_*^{(t-1)}$	Adjusted previous state memory gate
I	Identity matrix	$f_g^{(t)}$	Forget gate
$A_{i,j}$	Cosine similarity of the and districts	$i_g^{(t)}$	Input Gate
m	Number of samples of data	$o_g^{(t)}$	
E_t	Yearly embedding output of mixhop GCN	$\tilde{c}^{(t)}$	
d	District name	c^t	
σ	Activation function	h^t	
		r	