Deep Learning Models for Satellite Image Time Series Analytics for Earth Observation Applications

THESIS

Submitted in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

by

Arshveer Kaur

ID No. 2017PHXF0432P

Under the Supervision of

Prof. Navneet Goyal

Senior Professor & HOD, Department of Computer Science and Information Systems

Birla Institute of Technology and Science, Pilani, Pilani Campus, India

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

Pilani Campus, Rajasthan, India

September, 2024

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

PILANI CAMPUS, RAJASTHAN, INDIA

CERTIFICATE

This is to certify that the thesis entitled, **"Deep Learning Models for Satellite Image Time Series Analytics for Earth Observation Applications"** and submitted by **Ms. Arshveer Kaur** ID No. **2017PHXF0432P** for the award of Ph.D. Degree of the institute embodies original work done by her under my supervision.

_____

Signature of the Supervisor

Name : **Prof. Navneet Goyal**

Designation : **Senior Professor, Department of CSIS, BITS**

Pilani, Pilani Campus, India

Date: $25/September/2024$

Dedicated to

My Family

ACKNOWLEDGEMENTS

I would like to express my gratitude to several personalities for their constant guidance, help, support, and well-wishes. I would like to thank Professor V Ramgopal Rao, Vice Chancellor, BITS Pilani, Prof. Sudhir Kumar Barai, Director, BITS Pilani, and Prof. Shamik Chakraborty, Associate Dean, Academic Graduate Studies and Research Division (AGSRD) for giving a conducive research environment to our Institute, without which my Ph.D. would not have been possible. I would next like to thank Prof. Navneet Goyal (Head, Department of CSIS) for providing the required research facilities to our department. I would like to acknowledge the support received from Dr. Shashank Gupta, DRC Convenor, in their timely help in streamlining the review process and completing all the formalities smoothly.

I express my sincere gratitude to my supervisor, Prof. Navneet Goyal, whose guidance and unwavering confidence have been invaluable throughout this journey. His encouragement served as a constant source of motivation, pushing me to strive for excellence and persevere through challenges. I am deeply grateful for his guidance, profound insights, and clear vision, which has enriched my academic experience but also contributed significantly in shaping the direction and success of this work.

In addition to my supervisor, I would also like to extend my heartfelt gratitude to Prof. Poonam Goyal for her invaluable contributions, patience, insightful discussions, and unwavering guidance in my work. Her willingness to engage in fruitful discussions and provide constructive feedback at every step of the process has greatly enriched the quality of my research. Moreover, I am deeply thankful to Prof. Poonam Goyal for her role in nurturing not only my professional development but also fostering my personal growth. Her mentorship has been a guiding light, and I am truly grateful for the opportunity to learn from her expertise and wisdom. Her constant support has not only enriched my academic pursuits but also shaped me into a more capable and confident individual.

I am deeply grateful to my Doctoral Advisory Committee (DAC) members, Dr. Sundaresan Raman, and Dr. Kamlesh Tiwari from BITS-Pilani, Pilani campus, for reviewing my research proposal as well as this thesis, including their valuable suggestions throughout the research work.

I would like to express my heartfelt appreciation to all the members associated with Advanced Data Analytics and Parallel Technologies (ADAPT) Lab, Neha and Anupama for continually helping me academically as well as personally and inducing a warm atmosphere in our labs.

Words fail to express my gratitude to my parents S. Amrik Singh and Mrs. Varinder Kaur for their love, motivation, and blessings. I am grateful to my brother Gurparkash Singh for his love, affection, and constant support both academically and personally. I

am truly grateful for his presence in my life, which has brought warmth, companionship, and inspiration. It was my family's constant encouragement, love, and belief in me that kept my determination to succeed. Words cannot convey the depth of my gratitude for their endless sacrifices and relentless motivation, which have shaped me into the person I am today.

I want to express my gratitude to my dear friends Stuti, Shail, Pratibha, Swamy, Ashok, Teja, and Vishwjeet for their efforts to make our journey delightful, and a memorable one. Each of them has contributed in their own unique way, be it with words of encouragement during challenging times, sharing joyous celebrations of success, or simply being there as a pillar of strength and understanding. Their efforts not only lightened the burden of academic rigors but also added layers of warmth and friendship that have made our collective journey truly unforgettable. It was their unwavering support and friendship that made every milestone sweeter and every hurdle easier to overcome.

A special thanks to my Fiance, Karansher Singh. Even though we met towards the end of this journey, his unwavering support and love lifted me through the toughest times when I missed my friends the most. His presence alleviated my loneliness, and his encouragement inspired me to persevere through every obstacle. Thank you for being my partner, and my source of strength.

Above all, the infinite grace of the Almighty GOD is of essential importance, and I solemnly offer my regards for his grace which enabled peace and harmony for this work.

Arshveer Kaur

ABSTRACT

Satellite data has emerged as a disruptive technology in the field of Earth observation, offering unparalleled insights into environmental dynamics and facilitating informed decision-making in various domains such as agriculture, land use management, and disaster monitoring. This thesis undertakes a thorough exploration of satellite data, contextualizing its importance alongside traditional numeric datasets. Through meticulous analysis and empirical research, the complexities inherent in satellite data acquisition are examined, with a keen focus on spatial, temporal, and spectral resolutions offered by different satellite platforms.

The thesis proposes two models for crop yield prediction namely YieldPredictNet (YPN), CropYieldNet (CYN). YPN and CYN leverage cutting-edge deep learning architectures to integrate multi-modal data sources, including meteorological, soil, and satellite-derived features, thereby enabling precise timely, and early crop yield prediction. Notably, YPN and CYN incorporate novel attribute selection and depth selection modules to optimize feature representation and mitigate noise inherent in the data. Additionally, YPN is equipped with a spatial clustering technique and temporal padding mechanism to capture spatial and temporal patterns inherently present in the agricultural domain. YPN recommends that taking data at week granularity predicts the yield more accurately by approx 3%, 8.5%, and 23% for soybean, wheat, and corn, respectively than taking the data at month granularity. Also, there is approx. 1.17%, 20.40%, and 11.12% improvement in respective yield predictions by modeling depth-variant factors. Data augmentation used in CYN showed maximum improvement of 23.2% for Sentinel-2, followed by Landsat-8 with 22.09%, and the least in MODIS with 18.99%.

Working with high spatial resolution satellite images is a challenging task due to their large computing requirements. PatchNet, another model proposed in the thesis helps in democratizing the use of satellite image technology for various earth observation applications. It enables efficient processing of high-resolution satellite image time series

by innovatively combining beam search and attention mechanisms to select and process the most informative image patches, thereby circumventing computational bottlenecks associated with high spatial resolution satellite image time series. PatchNet achieves state-of-the-art performance for various Earth observation applications considered. Patch Selection Mechanism has a significant improvement in the model performance as RMSE achieved by random selection is 24.29 bu/ac and 9.98 bu/ac for corn and soybean yield prediction in comparison to 21.47 bu/ac and 7.29 bu/ac, respectively using PatchNet.

Furthermore, the thesis addresses the complex problem of trade-off between spatial and temporal resolutions in satellite systems through two data fusion techniques—LSFuseNet and FuSITSNet. LSFuseNet works on histogram time series of satellite data whereas FuSITSNet works on image time series. Both these deep learning models employ sophisticated fusion mechanisms to seamlessly integrate data from satellites with varying spatial and temporal resolutions, without significantly increasing computational overhead. This helps in unlocking new opportunities for downstream analysis and interpretation.

Spectral Reflectance Indices (SRIs) represent an important data modality in satellite systems. But, SRIs have not been extensively explored. The thesis proposes SpInN, an automated model for spectral reflectance index selection, which innovatively combines video (ViViT) and text (BERT) transformers to recommend relevant spectral indices for a specific Earth observation application. SRIs are typically associated with loss of spatial information. To overcome this limitation, we introduce the concept of an SRI image.

Lastly, we propose a foundation model SaTran for learning end task agnostic representation of Satellite Image Time Series which otherwise have huge computational requirements. SaTran focuses on non-redundant patch tubes using its two-fold mechanism for handling redundancy and distributed application of VideoMAE to enable space and time-efficient processing of Large-size SITS.

Through rigorous experimentation, comparative analysis, ablation study, and innovative methodologies, the thesis proposes to significantly advance the broad area of Artificial Intelligence for Earth Observation (AI4EO) by bringing together two disruptive technologies - Deep Learning and Satellite Imagery.

Contents

List of Abbreviations

AI :	Artificial Intelligence
AI4EO :	Artificial Intelligence for Earth Observation
ASU :	Attribute Selection Unit
AVHRR :	Advanced Very High-Resolution Radiometer
BERT :	Bidirectional Encoder Representations from Transformers
CC :	Crop Cycle
CCP :	Cloud Cover Prediction
CMA :	Cross-Modal attention
CNN :	Convolutional Neural Networks
CTM :	Core Temporal Module
CYN :	CropYieldNet
CYP :	Crop Yield Prediction
DA :	Data Augmentation
DER :	Dual Encoder and Recommender
DLSU :	Depth-level Selection Unit
DNN :	Deep Neural Networks
DSM :	Depth-level Selection Module
DTM :	Downstream Task Module
DTW :	Dynamic Time Warping
EG :	Embedding Generation
EO :	Earth Observation
ESA :	European Space Agency
EVI :	Enhanced Vegetation Index
FAM :	Feature Alignment Module
FAS :	Flat Attribute Selection

FM : Fusion Module

GAN : Generative Adversarial Network

GATR : Global Automated Target Recognition System

GEE : Google Earth Engine

GELU : Gaussian Error Linear Unit

GNDVI : Green Normalized Difference Vegetation Index

GRU : Gated Recurrent Unit

HLS : Harmonized Landsat Sentinel-2

KNN : k-nearest neighbor

LAI : Leaf Area Index

LASSO : Least Absolute Shrinkage and Selection Operator

LST : Land Surface Temperature

LSTM : Long Short-Term Memory

MAE : Mean Absolute Error

MHP : Multi-time horizon prediction

ML : Machine Learning

MLP : Multi-layer Perceptron

MODIS : Moderate Resolution Imaging Spectroradiometer

MSAVI : Modified Soil-Adjusted Vegetation Index

MSE : Mean Squared Error

MSR : Modified Soil Ratio

MSTE : Multispectral Spatiotemporal Encoders

NASA : National Aeronautics and Space Administration

NDSI : Normalized Difference Snow Index

NDVI : Normalized Difference Vegetation Index

NDWI : Normalized Difference Water Index

NDYI : Normalized Difference Yellowness Index

NIR : Near Infra-Red

NLP : Natural Language Processing

NOAA : National Oceanic and Atmospheric Administration

NS : Neighbor Selector

PAM : Patch Alignment Module

PCC : Padded Crop Cycle

PSM : Patch Selection Module

PSRI : Plant Senescence Reflectance Index

RBF : Radial Basis Function

ReLU : Rectified Linear Unit

RF : Random Forest

RMSE : Root Mean Square Error

RNN : Recurrent Neural Networks

RSFN : Robust Spatiotemporal Fusion Network

SAM : Spatial attention mask

SAR : Synthetic Aperture Radar

SAVI : Soil Adjusted Vegetation Index

SCP : Snow cover Prediction

SD : Sustainable Development

SDE : Soil Data Encoder

SDGs : Sustainable Development Goals

SDM : Soil Depth Modelling

SEP : Solar Energy Prediction

SGD : Stochastic Gradient Descent

SHP : Single-time horizon prediction

SITS : Satellite Image Time Series

SMAP : Soil Moisture Active Passive

SMP : Soil Moisture Prediction

SNN : Spike neural network

SpInN : Spectral Index Network

SPOT : Satellite Pour l'Observation de la Terre

SPS : Seasonal Prediction System

SR : Simple Ratio

SRE : Surface Reflectance Encoder

SRIs : Spectral Reflectance Indices

STARFM : Spatial and Temporal Adaptive Reflectance Fusion Model

SVM : Support Vector Machine

SVR : Support Vector Regression

SWIR : ShortWave Infrared

TSE : Time Series Encoder

TSM : Task-Specific Module

UAV : Unmanned Aerial Vehicles

USDA : United States Department of Agriculture

ViT : Vision Transformer

ViViT : Video Vision transformers

WDRVI : Wide Dynamic Range Vegetation Index

WOFOST : World Food Studies

Y : Yearly

YPN : YieldPredictNet

List of Tables

List of Figures

Chapter 1

Introduction

Artificial intelligence (AI) and satellite imaging technologies are two disruptive technologies which we attempt to bring together in this thesis. With the ability to process vast amounts of data and recognize intricate patterns, AI has demonstrated its utility across numerous domains, including healthcare, finance, and transportation. Satellite Imaging Technologies, on the other hand, encompass a range of technologies used to capture high-resolution images of the Earth's surface from orbiting satellites. These images provide invaluable insights into various aspects of the planet, such as land use, vegetation cover, urban development, and environmental changes over time. By combining the analytical power of AI with the information provided by satellite imagery, we seek to unlock new avenues for understanding and addressing complex earth observation applications such as prediction of crop yield, soil moisture, solar energy, cloud cover, snow cover, etc.

1.1 Evolution of Satellite Imaging Techniques

The evolution of satellite image technology has been characterized by ongoing advancements across various dimensions, including improvements in sensor technology, spatial resolution, spectral capabilities, revisit frequency, and data processing techniques.

The first Earth observation satellite, launched in the early 1960s is the TIROS-N series which provided low-resolution images of the Earth's surface, primarily for weather monitoring [9]. The next launched satellite is one of longest-running satellites "the Landsat Program", a series of Earth-observing satellite missions jointly managed by National Aeronautics and Space Administration (NASA) and the U.S. Geological Survey. The first satellite of the series is Landsat-1 which was launched in 1972, and currently, the actively orbiting satellites are Landsat 8, and Landsat 9.

The SPOT (Satellite Pour l'Observation de la Terre) Program was launched in 1986 for commercial purposes. Synthetic Aperture Radar (SAR) technology, which allows for day-and-night and all-weather imaging, was integrated into satellite systems and function from the 1980s to the 1990s.

Another successful satellite used across the world is the Moderate Resolution Imaging Spectroradiometer (MODIS) [10]. First MODIS instrument was the Terra satellite launched in December 1999, and the second was Aqua launched in May 2002. MODIS has a viewing swath width of 2,330 km and a revisit time ranging from every one to two days to 16 days, depending upon the product. Its detectors measure 36 spectral bands and it acquires data at three spatial resolutions: 250 m, 500 m, and 1,000 m [11].

Another commonly used satellite is the Sentinel satellite. It is a series of Earth observation satellites developed by the European Space Agency (ESA) as part of the European Union's Copernicus program [12]. Planet is a commercial satellite working with 200 satellites that together provide an unprecedented dataset of Earth observation imagery. The data provided by these satellites is high-resolution images, but it is not available free of cost [13]. The details of popular satellites are explained in the sub-sections below.

1.1.1 Advanced very high-resolution radiometer

Advanced very high-resolution radiometer (AVHRR) is the TIROS-N series satellite which has been used in the earlier days for different Earth Observation applications. AVHRR was launched in 1978 by NASA. It is carried by National Oceanic and Atmospheric Administration (NOAA) series of satellites and is mainly used for weather surveillance, sea

surface temperature, and detection of wildfires. It provides data for five channels including red, near-infrared, and two thermal radiation bands [9]. The satellite provides data at a spatial resolution of 1 km and a temporal resolution of 1 day.

AVHRR suffered from certain limitations, one of which is the coarse spatial resolution. Additionally, the data is captured for a limited number of spectral bands, which is insufficient to discriminate between different land cover types and accurately analyze the earth's surface. The satellite does not have advanced onboard data processing capabilities, which delays the availability of processed data. The users have to process the data on their own which is a challenging process. Moreover, the satellite data is potentially affected by atmospheric conditions such as clouds and aerosols.

1.1.2 Satellite Pour l'Observation de la Terre

The SPOT (Satellite Pour l'Observation de la Terre) satellites are a series of Earth observation satellites operated by Airbus Defence and Space. The SPOT program is one of the longest-running commercial Earth observation satellite programs and has contributed significantly to applications such as cartography, agriculture, forestry, urban planning, and environmental monitoring [14]. The recent launches of the SPOT program are SPOT-6 and SPOT-7 which provide data at a high spatial resolution of less than 1.5 m for optical bands and 6 m for multispectral bands including red, blue, green, and near-infrared. The satellite has not been used in any of the applications in the literature due to the unavailability of data to the public.

1.1.3 Moderate Resolution Imaging Spectroradiometer (MODIS)

Moderate Resolution Imaging Spectroradiometer (MODIS) is the most commonly used satellite capturing data using two sub-satellites - Terra and Aqua. Terra passes from north to south across the equator in the morning, while Aqua passes south to north over the equator in the afternoon. They view the entire Earth's surface every day or every 2 days acquiring data in 36 spectral bands. These bands are acquired using various sensors (called data products).

Table 1.1: Products of MODIS data

Data Product	Description	Resolution
MOD09A1	Contains 7 surface reflectance bands and other quality check bands	Spatial – 500 m Temporal – 8 day
MOD09Q1	Contains same bands as in MOD09A1	Spatial – 250 m Temporal – 8 day
MOD09GA	Contains same bands as in MOD09A1	Spatial – 500 m Temporal – daily
MOD11A1	Provides Land Surface Temperature (LST) at day and night time	Spatial – 1 km Temporal – daily
MOD11A2	Contains same bands as in MOD11A1	Spatial – 1 km Temporal – 8 day
MCD12C1	Gives yearly description of land cover type	Spatial – 500 m Temporal – yearly
MOD13A1	Provides Vegetation Index values Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) along with other reflectance bands	Spatial – 500 m Temporal – 16 day

There are numerous data products with different temporal and spatial resolutions used for capturing different information. Table 1.1 presents a few data products describing the type of bands they capture. The most commonly used and suitable data products for agriculture-related study include MOD09, MOD11, MCD12, etc.

1.1.4 Landsat

NASA and U.S. Geological Survey jointly manage a series of Earth-observing satellite missions under The Landsat Program. Landsat satellites have good ground resolution and spectral bands to effectively track land use and land change. Land use changes [15, 16] due to climate change, urbanization [17–20], wildfire [21], biomass changes, etc.

The Landsat Program is the longest-running program for capturing Earth's satellite imagery. The first satellite "Landsat 1" was launched in 1972. After that other satellites were launched and terminated at various time intervals. The currently active satellite Landsat 8 was launched in 2013. The timeline for the satellite program is given in Figure 1.1.

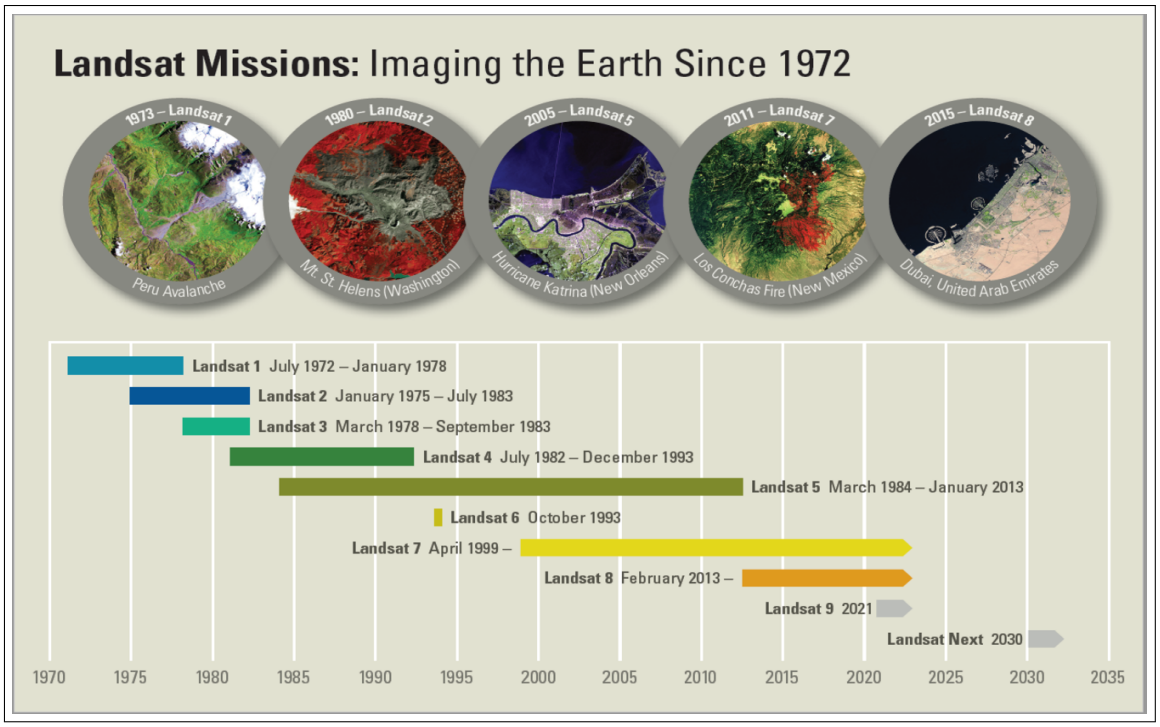

Figure 1.1: Landsat Timeline [1]

1.1.5 Sentinel

Sentinel-1 is the first Copernicus satellite consisting of two satellites sharing the same orbital plane. Sentinel 1 has a spatial resolution of 5m and a repeat time of 12 days. Sentinel-1 provides radar imagery for a wide range of applications. Synthetic Aperture Radar (SAR) images are highly effective for monitoring land subsidence and structural damage, as their systematic observations and advanced interferometric capabilities can detect ground movements that are almost imperceptible in daily life. This data is not only valuable for urban planners but also crucial for tracking changes caused by earthquakes, landslides, and volcanic activity. Additionally, it supports geohazard monitoring, mining, geology, and city planning by assessing subsidence risks. Sentinel-1 is specifically designed to deliver rapid-response imagery for disasters like floods and earthquakes.

Another satellite Sentinel -2A was launched in 2015. It delivers an exceptional view of Earth through its combination of high-resolution imagery, innovative spectral capabilities, a 290 km swath width, and frequent revisit intervals. The primary objective of the

Table 1.2: Resolution of various satellite systems

Satellite System	Spatial Resolution	Temporal Resolution	No. of bands
AVHRR	1km	Daily	4
MODIS	500m	8 days	7
SPOT	20m	3	5
Landsat-7	30m	16	8
Landsat-8	30m	16	11
Sentinel-2	20m	10	13
RapidEye	5m	Daily	5

mission is to provide critical data for agricultural and forestry management, thereby supporting food security. The satellite imagery facilitates the calculation of various plant indices, including leaf area, chlorophyll content, and water content, which are vital for precise yield forecasting and vegetation monitoring. Beyond tracking plant growth, Sentinel-2 is used in mapping land cover changes and monitoring global forest ecosystems.

1.1.6 Planet

Planet is a private Earth imaging company that operates a large constellation of small satellites dedicated to capturing high-resolution imagery of the Earth's surface [13]. It consists of several small satellites called CubeSats, and different generations of satellites, such as the Dove and SkySat series. The Dove satellites are the primary imaging satellites that capture data in the visible and near-infrared spectral ranges. SkySat series is a collection of larger satellites equipped with higher-resolution optical and synthetic aperture radar (SAR) sensors. The satellites provide data at high spatial and temporal resolutions but a lesser number of bands in comparison to MODIS, Landsat, and Sentinel. Moreover, the data is not publicly available which is one of the main reasons these satellites are not much used.

1.1.7 Resolutions of Satellite Systems

We now give the details of resolutions of the most popular satellite systems in a tabular form in Table 1.2.

Satellite systems face a trade-off between spatial and temporal resolutions, making it challenging to optimize both simultaneously. Sensors with high spatial resolution often cover a smaller area compared to those with lower spatial resolution. Consequently, a smaller field of view leads to longer time requirements to survey the same area. Thus, as spatial resolution increases, temporal resolution decreases. Freely available imagery, such as Landsat, Sentinel, and MODIS, typically offers either a short revisit time measured in days (1-4 days) with resolutions ranging from 300m to 500m, or a longer revisit time measured in weeks (10-20 days) with resolutions ranging from 10m to 30m.

1.2 Importance of Satellite Imaging Technology

The increased application of satellite imagery in environmental forecasting, border area surveillance, security of energy resources, mapping construction, and swift responses to emergencies, like natural disasters, and defense security concerns, is driving the escalating demand for satellite imaging [22].

The significance of enhanced satellite system technology is increasingly evident across various domains, contributing substantially to driving the growth of the satellite imaging market value. As per the reports [2] and graph shown in Figure 1.2, the Satellite Data Services Market Size reached USD 198.9 billion in 2022. The industry is anticipated to witness significant growth, projecting an increase from USD 128.359 billion in 2023 to USD 246.9133 billion by 2030.

The growing integration of artificial intelligence (AI) and machine learning (ML) in the space sector has created a plethora of opportunities for researchers and scientists working in different domains all across the world. ML algorithms play an increasingly vital role in processing daily satellite imagery, enabling the classification and detection of objects, identification of geographic and topographic features, and monitoring subtle changes over time. Defense corporations have introduced the Global Automated Target Recognition System (GATR) specifically designed for recognizing satellite pictures. GATR employs open-source deep learning libraries to efficiently classify and identify extensive datasets in a faster and more effective manner. These advancements are poised to

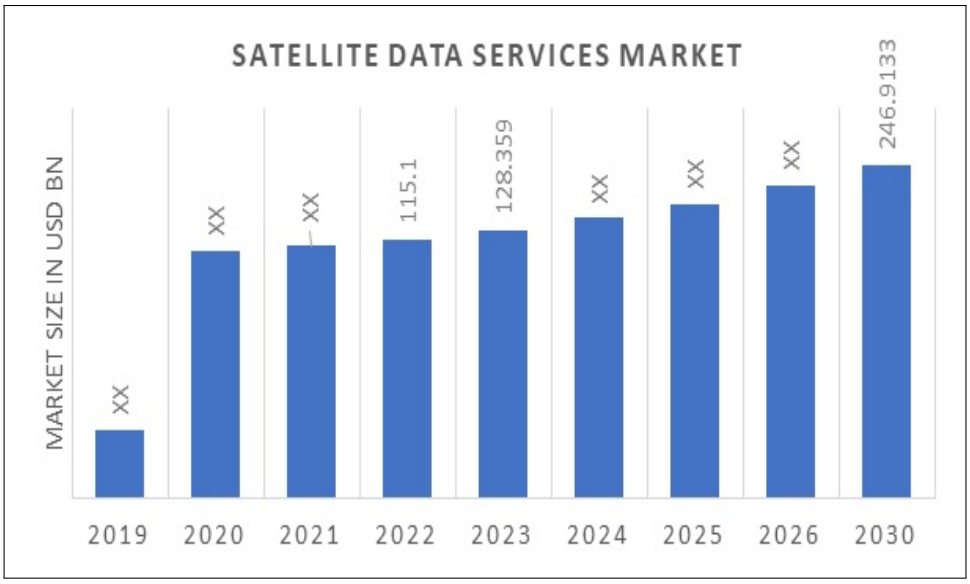

Figure 1.2: Anticipated Market value trend for Satellite Imagery [2]

create opportunities for the satellite data services market in the near future. Furthermore, ML and AI models demonstrate accurate detection capabilities for various elements visible from space, including cars in parking lots, crop yields, and other Earth Observation applications. .

1.3 Artificial Intelligence for Earth Observation (AI4EO)

Artificial Intelligence for Earth Observation (AI4EO) involves the integration of AI techniques with the vast amount of data generated by Earth observation satellites. Integrating Earth Observation (EO) and Artificial Intelligence (AI) technologies is pivotal in addressing climate change's impacts, enhancing disaster management operations, and solving problems in other areas like transportation, urban planning, agriculture, changes in land cover, deforestation rates, and water resource availability, etc. The application of AI models to analyze this data facilitates the development of intelligent prediction models applicable to different domains.

As per the United Nations, leveraging Earth Observation (EO) data, such as satellite images, proves advantageous for generating and supporting official statistics, serving as a valuable complement to traditional sources of socio-economic and environmental data.

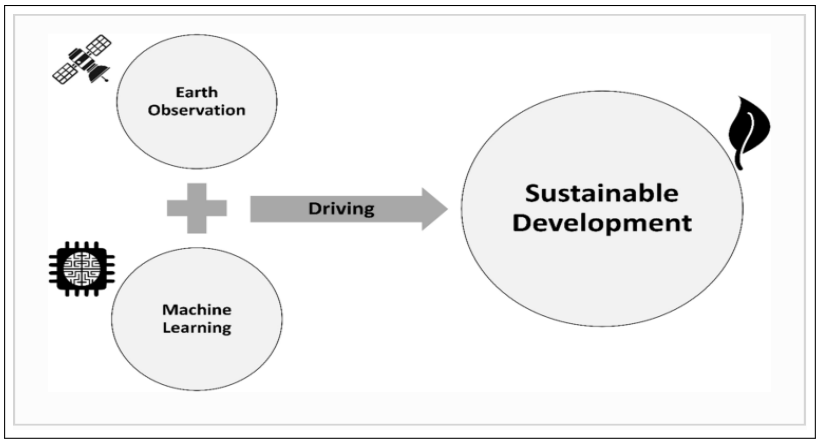

Figure 1.3: AI4EO leads to sustainable development [3]

Satellite imagery emerges as a cost-effective technology, particularly capable of providing data at a global scale. This global accessibility becomes pivotal for understanding the progress and contributions of underdeveloped countries toward Sustainable Development (SD), considering their limited resources for data collection. However, substantial volume of data furnished by EO sources necessitates effective analysis and processing through appropriate methods and tools to yield robust indicators of SD. The continual growth of the Machine Learning (ML) field presents new opportunities for monitoring and analysis of satellite images applied to Sustainable Development Goals (SDGs). The intersection of EO data and ML methodologies enhances the precision and efficiency of deriving meaningful insights relevant to SDGs [3] (Figure 1.3).

AI has resembled the “electricity of the 21st century” and has transformed the world. Among the various branches of AI, machine learning plays a pivotal role by bridging the gap between the ever-expanding, often openly accessible data and the development of solutions and products derived from that data. Earth observation community stands to greatly benefit from the integration of AI [23].

1.4 Foundational Deep Learning Models

Deep learning is the subset of Artificial Intelligence (Figure 1.4) and has proven to be the game changer in the era of AI. As a cutting-edge technology, deep learning finds applications across a broad spectrum of fields.

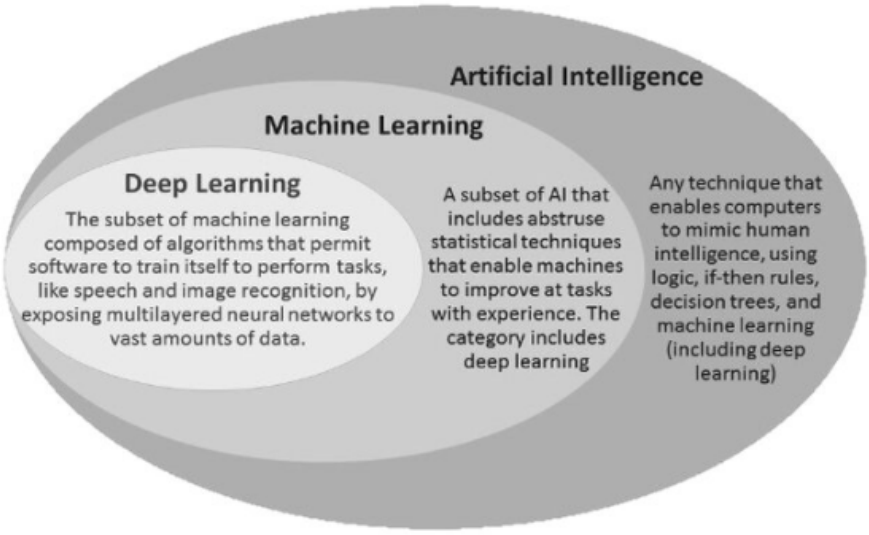

Figure 1.4: AL-ML-DL Relationship [4]

Over the past years, deep learning has undergone significant advancements across various models, each tailored to address specific challenges and tasks. In this thesis, we have used many deep learning models individually or in combination with one another. We have now given a generic brief description of the models used in our work.

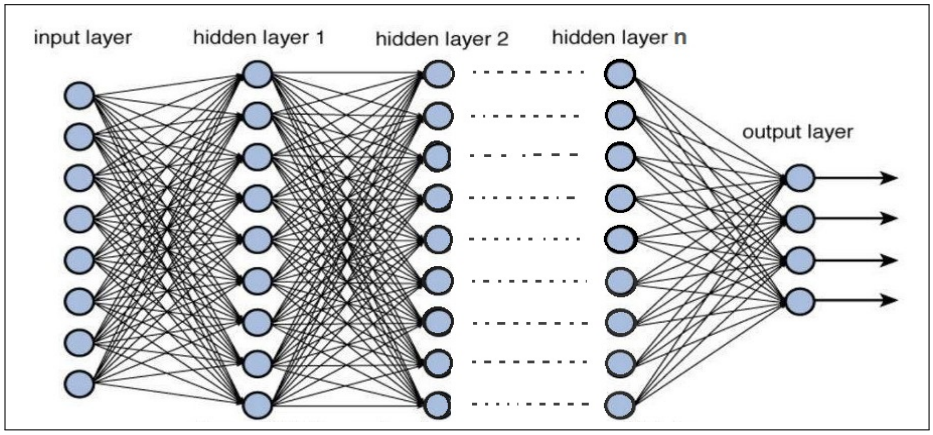

Figure 1.5: Deep Neural Network [5]

Starting with the fundamental **Deep Neural Networks (DNNs)**. These networks, characterized by multiple layers of interconnected nodes, have seen continual refinement in optimization algorithms and training methodologies, enhancing their ability to learn intricate patterns from data (Figure 1.5).

Convolutional Neural Networks (CNNs) have emerged as powerhouse models for tasks involving images, leveraging hierarchical feature extraction to discern complex visual patterns. This makes CNNs highly effective in computer vision applications, such as image classification and object detection (Figure 1.6).

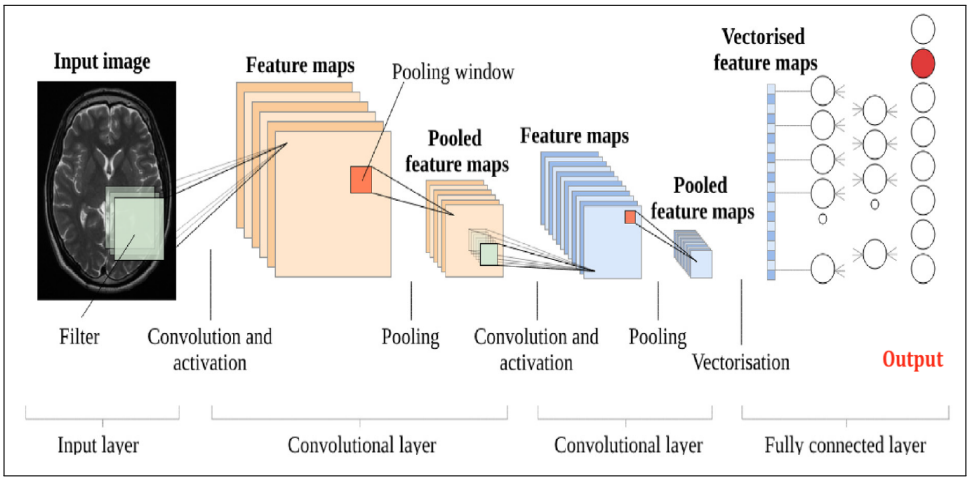

Figure 1.6: Convolutional Neural Network [6]

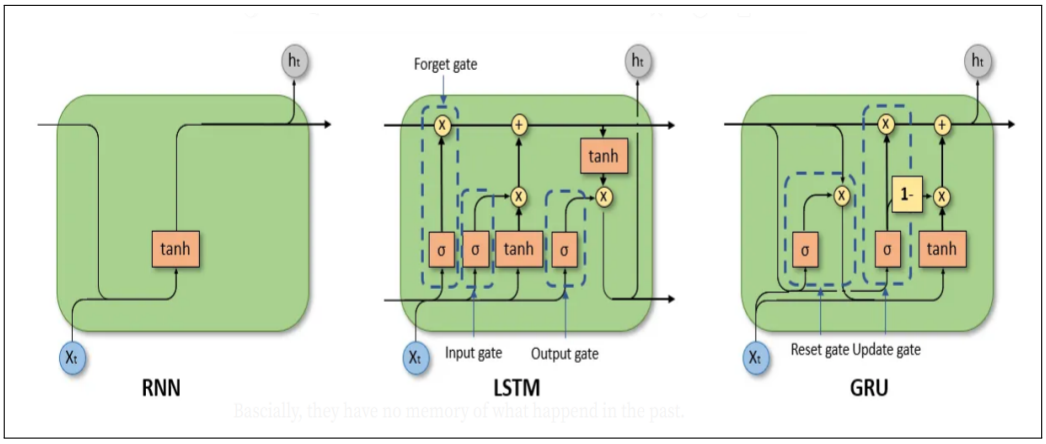

Figure 1.7: Recurrent Neural Network and its variants [7]

Recurrent Neural Networks (RNNs), with variants like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), specialize in sequential data processing, addressing challenges like vanishing gradients and allowing models to capture and remember dependencies over extended sequences (Figure 1.7). This makes them well-suited for tasks such as natural language processing and time-series analysis.

Moving to advanced models, **Autoencoders** (Figure 1.8), a class of unsupervised learning models have gained prominence for their ability to compress data and extract meaningful features. They consist of an encoder, compressing input data into a latent space, and a decoder, reconstructing the input from this compressed representation. They find applications in dimensionality reduction and anomaly detection where the focus is on capturing the essential features of input data. It makes them a valuable tool for representation learning.

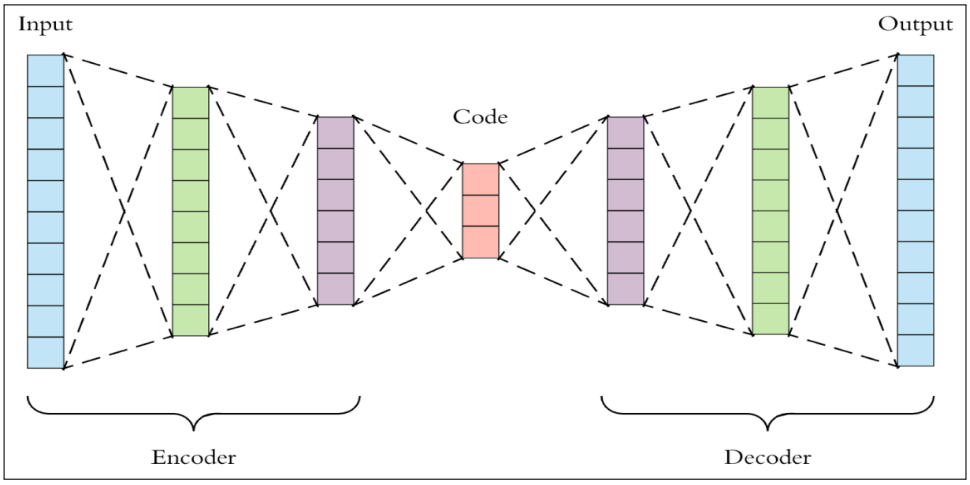

Figure 1.8: Autoencoder [8]

The **Transformer** architecture, initially designed for natural language processing, has emerged as a versatile model architecture. Their attention mechanisms enable capturing contextual relationships in data, making them effective for tasks beyond sequential data. Transformers have been successfully applied in computer vision, where they outperform traditional architectures, providing enhanced performance in image recognition and object detection. Their ability to process data in parallel and capture long-range dependencies makes them particularly valuable in scenarios where contextual understanding is crucial. Furthermore, Transformers have been pivotal in advancing transfer learning approaches. Models pre-trained on large datasets, such as Bidirectional Encoder Representations from Transformers (BERT) for language understanding or Vision Transformer (ViT) for vision tasks, have demonstrated remarkable performance when fine-tuned on

smaller, task-specific datasets. This has led to breakthroughs in various domains by leveraging the knowledge encoded in pre-trained models.

Together, these models and methodologies showcase the great progress in deep learning and we have used them in this thesis for solving the earth observation applications.

1.5 Motivating Earth Observation Applications

Proper modeling of Earth Observation Applications is critically important for developing acceptable and deployable solutions. Choosing spatial and temporal granularities is done in an application-aware manner and is dependent on multiple factors like availability of computing resources, availability of ground truth, real-timeliness of application, scalability requirements, etc. A detailed explanation of the choice of suitable granularities is given next for different earth observation applications considered in the thesis.

Crop Yield Prediction: By 2050, the world population is expected to touch the 10 billion mark. At the same time, global warming and climate change are playing havoc with the weather patterns globally and their impact on crop yield needs to be analyzed very carefully. Moreover, increasing unplanned urbanization is shrinking cultivable land, especially in India. Feeding an additional 2.5 billion people in the next three decades is going to be a herculean task despite the technological advances and innovations in the agricultural sector in the 21st century. Predicting Crop yield is a crucial aspect of agriculture that has garnered significant attention from scientists due to its profound influence on the national and international economy, as well as its potential to address the issue of food scarcity.

Accurate and timely crop yield forecasting plays a crucial role in supporting governments and farmers in making informed decisions and planning strategies related to agricultural production. Governments rely on crop yield prediction to formulate policies regarding import and export, determine pricing mechanisms, and address potential threats to food security. By having access to reliable crop yield predictions, governments can effectively manage resources, allocate budgets, and implement measures to ensure stability and sustainability in the agricultural sector. For individual farmers, crop yield predic-

tion provides valuable insights into the expected output of their crops. This information enables them to plan their farming activities, make informed choices regarding inputs such as seeds, fertilizers, and pesticides, and manage their resources effectively. Farmers can adjust their cultivation practices, adopt appropriate risk management strategies, and optimize their harvest and post-harvest activities based on the anticipated yield.

Crop yield prediction is required at the temporal granularity of a crop cycle which is typically 6-8 months, and ideally at the farm-level spatial granularity. While crop cycle temporal granularity is achievable, working at farm-level granularity poses a lot of challenges. These include lack of ground truth at the farm level, increased computational requirements of working high spatial resolution satellite images, farm segmentation due to irregular shape, etc. Predicting crop yield at crop cycle granularity enables yearly prediction which is of paramount importance for governments as it allows for strategic planning, ensuring a stable food supply through proper management of data warehouses and supply chains. In the thesis, we have worked with county/district level spatial granularity and at a crop-cycle temporal granularity. The predictions are made up to 8 weeks early from harvest. The predictions are updated at the revisit frequency of the satellite used. The prediction accuracy improves as we move closer to the harvest. We hypothesize that the counties experiencing 'similar' weather and soil conditions tend to have 'similar' yield patterns. Thus, we exploited the spatiality of the problem by clustering the counties. The model pre-trained on global data is fine-tuned with the data of individual clusters. In this way, the model can learn both global and cluster-specific yield patterns.

Soil Moisture Prediction: Another significant application is soil moisture prediction which indirectly impacts many other earth observation applications. It helps in studying long-term climate patterns, managing water resources in any area, drought management, and allocation of water resources for agriculture and urban use. It helps the infrastructure industry to decide on the design and material for bridges and pipelines for their stability and longevity. It even helps in disaster preparation by identifying high-risk zones for landslides or droughts. Soil moisture analysis helps environmental planners to implement erosion control measures effectively as less moisture in the soil leads to more soil erosion

in the area.

We predicted soil moisture at a spatial granularity of the county level and temporal granularity of a month. Soil moisture prediction is also governed by the revisit frequency of satellite data, update frequency of meteorological data, ground truth availability, and use case. Predicting soil moisture on a monthly basis helps in facilitating strategic planning for water availability, drought management, the allocation of resources for both agricultural and urban use, and infrastructure planning.

Solar Energy Prediction: Solar energy prediction is done to find a suitable location for the installation of solar plants and reduce the dependence on fossil fuels for economic development. It helps in planning resource utilization to meet energy demands. The predictions play a vital role in shaping the future of energy production and consumption, leading the way towards a more sustainable and eco-friendly energy source.

Solar energy can also be predicted at different granularities ranging from daily prediction to a few days or months. At different temporal granularities, it has a different significance. Predicting solar energy on a few days or monthly level helps governments to make policies to install solar infrastructure in solar energy-rich areas and make use of renewable energy sources for a sustainable environment. Considering the availability of ground truth and revisit frequency of satellite data, we predicted the average solar energy produced in a county at a fortnightly temporal granularity.

Snow Cover Prediction: Snow cover plays a vital role in agriculture as it provides moisture to the soil during spring melt. Snow cover prediction is crucial for managing reservoirs, planning irrigation, road maintenance, ensuring safer transportation, and preventing flooding. Also, predicting the snow cover can help in identifying high-risk areas to issue early warnings. It also helps in planning tourism for a nation during the winter season. Predicting snow cover can be carried out using drones, sensors, or by manually visiting the place. However, snow cover prediction with the help of satellite data is useful in getting an estimate of the presence of snow cover in dangerous and inaccessible areas.

Unlike CYP, snow cover prediction is required at a fine temporal granularity of a few hours to a month depending upon the availability of ground truth. It is typically governed

by the revisit frequency of satellite data and the update frequency of meteorological data. Keeping these factors in mind we have predicted snow cover area at a monthly granularity which allows us to study the impact of global warming and the influence of snowmelt on soil moisture which is particularly relevant for farmers and agricultural planners. The spatial granularity is at the county/district level.

Cloud Cover Prediction: Cloud cover prediction has a significant impact on the tourism industry and Maritime Operations of a nation. Tourism agencies and travelers consider historical cloud cover patterns when planning vacations and destinations. Mariners use cloud cover predictions, especially in coastal and navigational areas, to plan shipping routes. Cloud cover can affect visibility and safe navigation at sea. Cloud cover also helps in air quality monitoring as cloud cover affects the dispersion of pollutants in the atmosphere. Air quality researchers use cloud cover predictions to understand how pollutants disperse and accumulate, aiding in air quality monitoring and management. Hydrologists and water resource managers use cloud cover predictions to model evaporation rates from water bodies. Cloud cover influences the amount of solar radiation reaching the surface, affecting evaporation rates and water availability.

We have predicted cloud cover for a county at a fortnightly temporal granularity due to the availability of satellite data. Cloud cover prediction at a monthly level or a few days helps in making decisions at a big-picture level profoundly impacting the tourism industry, water resource managers, maritime navigation, etc.

All the above-listed problems are studied using satellite image time series for which we developed LSTM, CNN+LSTM, and other deep learning models. All the models work for time series data coming from satellites and other sources. Satellite data is used in different forms namely, histograms, images, and spectral reflectance indices.

1.6 Research Gaps

Many attempts have been made in the literature to solve earth observation applications using satellite data. Based on an extensive literature survey, we identified the following research gaps which we have addressed in the thesis:

1. *Impediment in the democratization of satellite imaging technology due to extensive computational requirements:* When working with high spatial resolution satellite image time series, the amount of data we need to process increases manifolds, leading to a computing bottleneck. Most of the work reported in the literature uses Spectral Reflectance Indices (SRIs) and histograms to deal with computing bottlenecks. But, this leads to the loss of critical spatial information. To the best of our knowledge, very little work is available in the literature that directly works with time series of raw satellite images for large spatial granularity prediction problems. Researchers have used satellite images for applications like object detection, semantic classification, etc. [24–26]

2. *Requirement for fusing data from 2 satellites:* The last decade has witnessed a significant improvement in sensor technology leading to the availability of higher spatial and temporal resolution satellite images. However, due to budgetary and technological constraints, it is not possible to capture satellite images with the required high spatial and temporal resolutions using a single satellite system. This necessitates the development of efficient fusion algorithms that combine high spatial resolution images of one satellite system (with low temporal resolution) with high temporal resolution images of another satellite system (with low spatial resolution). Many applications predicting crop yield, forest cover, forest fire, etc. require satellite image time series data at high resolution along both spatial and temporal dimensions. Publicly available data from satellite systems like LANDSAT 8/9, SENTINEL-2, MODIS, etc. have high resolution only along one dimension and not along the other dimension. For example, LANDSAT 8 has a spatial resolution of 30m and a 16-day revisit cycle whereas, MODIS has a spatial resolution of 250-500m and a temporal resolution of 8 days.

   To overcome this limitation, researchers have tried to fuse the data from satellite systems with complementary resolutions [27] and from different sensors of the same satellite [28–31] to produce high-spatial and spectral resolution imagery. In

most cases, models are developed to generate a synthetic image at a finer temporal resolution with the motivation that the combined data of original and synthetic images can be used for different earth observation applications. Because of the complex image generation process, these generative models operate on smaller-scale datasets, limited to only a few locations. In the realm of the applications at hand, where high-temporal-resolution time series data is imperative, the interpolation of images between consecutively captured images increases the data volume by at least twofold making it storage and computation intensive. Moreover, it potentially propagates existing noise in original images.

3. *Use of limited spectral reflectance indices:* Existing studies rely on a limited number (mostly only a single SRI is used and a maximum of 2-3) of SRIs manually picked by domain experts [32, 33]. These studies overlooked valuable information encoded in other SRIs which could be useful for a given application. Another major limitation of existing work involving SRIs is that a single SRI value is used to represent a large region, leading to loss of spatial information.

4. *Lack of foundation model:* Deep learning has gained popularity in the remote sensing community. A couple of studies used BERT [25, 26] to classify time series for every pixel which limits them from effectively exploiting the spatial correlations in the image time series. Moreover, these models work only for classification as they spatially segment the image time series which is not suitable for prediction problems like prediction of crop yield, snow cover, cloud cover, etc. In prediction problems, the ground truth is mostly available for coarser granularity than that of pixel e.g., at a county or a district level. Another model TSViT used ViT for landcover classification. The authors factorize input dimensions into spatial and temporal components to reduce the computation. However, the model is not able to identify the redundancy in the patches and processes them all.

1.7 Thesis Contributions

The summary of the thesis is presented below:

- *Establishing the advantages of using satellite data over numeric data:* We performed a preliminary study for crop yield prediction to compare the impact of ground-based data and satellite data. We have used a numeric NC94 dataset and satellite data from three satellites viz. MODIS, Landsat-8, and Sentinel-2. We propose a deep learning model **YieldPredictNet (YPN)** which works with numeric data and models the problem of crop yield prediction as a spatiotemporal problem. We proposed another model **CropYieldNet(CYN)** which uses a time series of histograms obtained from satellite images. We hypothesize and validate that high-resolution satellite data can provide better insights about the factors affecting crop yield, even when limited historical data is available for training suitably designed deep learning models.

- *Handling computational bottleneck and loss of spatial information:* To address the problem of infeasible computational requirements for processing high spatial resolution satellite image time series (SITS) we proposed a model **PatchNet** which learns prominent patterns in a SITS by doing a spatial patch-based partial traversal, e.g., $(1/p)th$ spatial processing of SITS using the idea of beam search and attention mechanism for learnable patch selection. The amount of processing is reduced by a factor of *p* with some additional overheads and the model still achieves state-of-the-art results for end tasks. Existing methods deal with the processing challenges by transforming the images into histograms which leads to loss of spatial information.

- *Handling spatial and temporal resolution trade-off:* We addressed the problem of resolution trade-off by developing fusion models for two pairs of satellites 1) For Landsat-8 and Sentinel-2 fusion and 2) for MODIS & Landsat-8 fusion. Both fusions have their own challenges which have been dealt with in this thesis.

  - For Landsat-8 and Sentinel-2, we proposed **LSFuseNet**, which fuses the histogram time series of the two satellites at the feature level. We chose to work with histograms because both satellites have a high spatial resolution, so working with a time series of huge-size images is computationally expensive. The proposed model LSFuseNet learns features from the individual time series of histograms from the two satellites with the help of respective pre-trained encoders and applies a novel dual-fusion using two modules viz. Fusion Module (FM) and Feature Alignment Module (FAM).
  - For MODIS-Landsat-2 Fusion, we proposed **FuSITSNet**, a twofold feature-based fusion model that can be used to fuse any two satellite image time series. FuSITSNet improves the temporal features of Landsat SITS by aligning its PatchNet processed patches with the MODIS SITS. It takes care of the untraversed area of the time series by cross-modality attention which assimilates complementary features from the two modalities (Landsat & MODIS).

- *Exploring a broader perspective of Spectral Reflectance Indices for different applications:* We proposed a generalized prediction model for different earth observation applications working on spectral indices as an input. We have listed 10 SRIs used in the literature for various purposes. We propose a model **Sp**ectral **In**dex **N**etwork (SpInN), which selects the most relevant spectral indices for a given application. SpInN performs dual encoding of SRI images at a timestamp using Video Vision transformers (ViViT) [34] (which we pre-train on different tasks and refer to as SRI ViViT). The dual pre-trained ViViT is fine-tuned using a disentangled representation learning in an end-to-end learning setup for downstream tasks. We applied BERT [35] to exploit temporal patterns in the obtained SRI time series.

- *Foundation Model:* SITS data can be characterized by the presence of patches with spatiotemporal redundancy persisting throughout the time series, referred to hereafter as redundant patch tubes. SITS data also contains patches where temporal redundancy lasts only for a few timestamps, referred to hereafter as non-redundant

patch tubes. We propose a transformer model, SaTran, for large size satellite image time series which exploits spatiotemporal redundancies. It has two modules - PatchTubeSelect and TemporalRedundancyHandler. We first remove spatiotemporal redundancies with the help of PatchTubeSelect which selects hotspots (non-redundant patch tubes) using an attention mechanism to discern critical areas necessitating focused attention and exclude the redundant patch tubes. We, then use TemporalRedundancyHandler which innovatively uses VideoMAE on non-redundant patch tubes to further handle temporal redundancy local to these patches.

1.8 Thesis Organisation

The thesis is organized in 8 chapters. Chapter 1 includes an introduction, and covers the background study, research gaps, and contributions of the thesis. The work carried out in this chapter helped us to identify the research gaps more precisely and to decide the future course of the thesis. Chapter 2 describes the datasets used and the pre-processing steps required for each type of dataset. Chapter 3 discusses the problem of crop yield prediction and the impact of using satellite data over conventionally collected data. In this chapter, we propose two models – YieldPredictNet for conventional data and CropYieldNet for satellite histogram time series. Chapter 4 proposes PatchNet for efficient representation learning of satellite image time series. Chapter 5 presents two fusion models – LSFuseNet and FuSITSNet to handle the mandatory trade-off between spatial and temporal resolution any satellite system faces. LSFuseNet and FuSITSNet fuse histogram time series and image time series, respectively taken from any two satellites. Chapter 6 presents the work SpInN a model working with satellite-obtained spectral reflectance indices which recommends SRIs relevant to an earth observation application. Chapter 7 introduces a foundation model SaTran for efficient processing large size SITS where the existing vision/video models shortfall. Chapter 8 concludes the thesis and throws an insight on future directions.

Chapter 2

Data Collection and Preparation

Earth observation involves the systematic collection, analysis, and interpretation of information related to Earth's physical, biological, and chemical systems using remotely sensed data collected using satellites, aircraft, and drones. The data collected is in the form of images from which various spectral reflectance indices (SRIs) can be derived. We have also worked with data collected using proximal sensors.

2.1 Sources of Data

The sources of data for earth observation applications are listed below:

1. *Ground-Based Data Collection:* Equipment is placed on or near Earth's surface to collect data for a specific purpose at a specific location. Different types of sensors are used to collect meteorological, environmental, and soil data. One major advantage of using Ground-based data collection is that temporal granularity can be adjusted as per the requirements of the application. Ground-based data collection has been prevalent but suffers from scalability problems and can be very expensive. The process of Ground-based data collection is also prone to human errors. Remotely sensed data captured using satellites overcome these problems. In Chapter 3, we have shown satellite data offers many advantages over ground-based data

to solve various earth observation applications. We have used NC94 [36] ground based data for crop yield prediction.

2. *Data collection through drones and aircraft:* Aerial photography is a technique of capturing images of the Earth's surface using Unmanned Aerial Vehicles (UAVs) and aircraft giving an aerial perspective of Earth's surface. They are typically equipped with different kinds of RGB, multispectral, and hyperspectral vision sensors which can be configured to take images at desired spatial and temporal resolutions. This method also suffers from scalability and cost issues. Any change in settling during flight introduces new challenges in analyzing data. E.g., any change in angle or height can result in misleading patterns which are difficult to interpret.

3. *Satellite Imaging:* Satellites have been in use for collecting data for the last five decades for various applications like communication, surveillance, earth observation, etc. In the past 5-6 years, satellite imaging technology has gained importance and adoption. This has been possible due to free access to data from different satellite agencies like NASA, ESA, ISRO, etc. This has led to the democratization of satellite imaging technology. Researchers across the world are now increasingly using satellite data to solve complex problems. Cloud service providers like AWS and Google Cloud are storing satellite data thereby further accelerating global usage of satellite imagery. Google Earth Engine (GEE) [37] and Microsoft's planetary computer [38] are leading platforms which provide application specific datasets collected from different satellite systems.

   The major advantage of satellite based remotely sensed data is that its coverage is global and is cost-effective as compared to ground-based and aerial data. It does not require any infrastructure on the ground to collect data and is therefore scalable. Satellite imaging technology is rapidly advancing in terms of spatial, temporal, and spectral resolutions. The finest spatial resolution available is 0.3 m using Worldview 3. Some satellite systems have a revisit frequency of 1 day. In terms of spectral resolution, we now can get hyperspectral images having 100s of

bands. Commonly used publicly available satellites have moderate resolutions for example, MODIS has a spatial resolution of 500m and a temporal resolution of 8 days, and Landsat-8 captures data at a spatial resolution of 30m at an interval of 16 days.

In this thesis, we have worked with ground-based numeric data and satellite data captured using MODIS, Landsat, and Sentinel satellite systems. The details of the datasets used in the thesis are given next.

2.2 Numeric Data: NC94

This dataset is collected by the North Central Regional Association of Agricultural Experiment Station for the North Central region of the United States and named as NC94 [36]. It is collected for 30 years from 1971 to 2000 at the county level and consists of crop data, soil data, and weather data. The dataset contains the yield data for almost all the major staple crops of the US including wheat, corn, sorghum, soybean, rice, etc. The crop data includes details about the harvested area for various crops, yearly production and yield of the crop for a county, yield unit, etc. The weather parameters collected in the NC94 dataset are maximum and minimum temperature, radiance, and precipitation. The maximum and minimum temperature is the highest and lowest temperature in the day for a location. Precipitation is the liquid or solid form of water that falls back to the ground. Radiance represents the light or heat emitted by the sun. Dew is the moisture condensed from the atmosphere in the form of small water drops on the crops and plants. Another type of data present in NC94 is the soil data which is collected once for the entire duration and is the same for a location for all considered years and thus is considered as static i.e. invariant with respect to time. The soil characteristics do not change in such a short period for a location. The dataset consists of 102 attributes of soil out of which 11 attributes are collected at 6 different depths (measured in cm) of the soil from the ground surface as shown in Figure 2.1. The description of the data attributes is given in Table 2.1.

Figure 2.1: Soil data at different depths

Table 2.1: NC94 Description

NC94 Dataset		
Crop Attributes	Crop name, area harvested, yearly production & yield of the crop for a county, yield unit	
Soil Attributes	Depth-invariant	PctArable, PctSlopeT, PAV, AcresT, Arable, PctSlopeA, Drainage, DepthH2O, DepthBed, OMkgm-2, MaxRoot
Soil Attributes	Varying with Depth	Sand, Silt, Clay, Liquid, Plastic, AvailH2o, BulkDen, Omatter, Perm, Wilt, Field
Meterological Attributes	maximum temperature, minimum temperature, Radiance, precipitation	
Additional Attributes		
Meteorological Attributes	Dew, Humidity, Visibility, Wind speed, Cloud cover	

Attributes such as sand, silt, clay, etc. vary with the depth of the ground and are referred to as depth-variant attributes. They impact yield prediction differently at different depth levels. Sand, Silt, and clay represent the percentage of these properties, respectively. Liquid and plastic is the presence of water content measured in centimeters (cm) at liquid and plastic limits. After a certain limit soil behaves like liquid and plastic. AvailH2o represents the water content available for the crop measured in cm. BulkDen is the bulk density measured as mg/m^3. Omatter is the percentage of organic matter present in the soil. Perm means permeability (cm/hr) which represents the property of transmit-

ting water and air through soil. The more permeable the soil is, the more it is easier to transmit water and nutrients to the roots of the crop passing through different layers of depth. The attributes like arable, depth Bed, drainage, etc are collected at a single depth and are thus referred to as depth-invariant.

Additional meteorological attributes: Since the NC94 dataset contains only 4 meteorological attributes, we used additional weather attributes viz. dew, humidity, visibility, wind speed, and cloud cover for all the counties for the same time duration of 30 years with the same granularity as that of NC94. These climate attributes have a significant effect on the crop yield. Dew activates the photosynthesis process in plants and crops by hydrating them. Humidity refers to the amount of water vapors present in the air. The right amount of humidity is an important factor in the growth of crops. The right amount of humidity helps in the growth and pollination process of the crops which directly affects the yield of a crop. The increased humidity can lead to the growth of bacteria and mold which reduces crop growth and sometimes destroys the crop. Visibility represents how far the objects can be seen clearly by the naked eye. The pollution particles in the air, snow, hale windblown dust, etc. are some of the reasons which reduce visibility and have an impact on the health of the crops. Thus, visibility helps in measuring the impact of many other factors affecting the crop yield. Wind speed is the speed with which wind flows from high to low-pressure areas. The turbulence caused by wind increases the carbon dioxide supply to the crops which increases the rate of photosynthesis and helps in the growth of crops. Cloud cover means the part of the sky covered by the clouds at a particular geographic location. Cloud cover reduces the temperature and radiance which can positively or negatively impact a yield depending on the crop requirements.

2.3 Satellite Data

These are satellites from NASA, European Space Agency, etc. End users can use platforms like Google Earth Engine to access, clean, and download this data by writing small scripts. Among these satellites, there are both free and paid options, each offering unique advantages and catering to different user needs and preferences. Many of these satellites

are operated by governmental agencies or international organizations and provide data without any cost e.g. MODIS, Landsat, sentinel, Advanced Very High-Resolution Radiometer (AVHRR), etc. Researchers, educators, policymakers, and practitioners leverage these datasets to gain insights into global phenomena, track long-term trends, and monitor changes in the Earth's surface over time. Moreover, the open nature of these datasets fosters collaboration and innovation within the scientific community, enabling the development of new methodologies, algorithms, and applications that benefit society as a whole.

In addition to free satellite data, there exists a range of paid options offered by commercial satellite operators and data providers. These premium datasets, such as RapidEye by Planet, Maxar Imagery Mosaics, and Airbus OneAtlas, often boast higher spatial resolution, increased temporal frequency, and specialized features tailored to specific industries or applications. While access to paid data typically involves subscription fees or licensing agreements, the benefits they offer can be invaluable for certain users.

The data used in this thesis is captured using three famous publicly available satellites viz. MODIS, Landsat-8, and Sentinel-2. A detailed description of each is given in the subsequent subsections.

2.3.1 MODIS

Moderate Resolution Imaging Spectroradiometer (MODIS) is the most commonly used satellite capturing data using two sub-satellites - Terra and Aqua. MODIS provides data through its various products listed in section 1.2.3. MODIS data used in this thesis is taken from MOD09A1. MOD09A1 has a spatial resolution of 500m and a revisit time of 8 days. The details of the bands of MOD09A1 are given in Table 2.2.

2.3.2 Landsat-8

The Landsat-8 data consists of 9 spectral bands with a spatial resolution of 30 meters and a temporal resolution of 16 days. The Landsat data is captured using two sensors – an operational land imager (OLI) with a spatial resolution of 30 meters and a Thermal Infrared

Table 2.2: Description of MODIS (MOD09A1) Product

Scientific Data sets	Description	Range	Scale Factor
Surface Reflectance band 1	Red Band	-100 - 16000	0.0001
Surface Reflectance band 2	Near Infra-Red (NIR) band	-100 - 16000	0.0001
Surface Reflectance band 3	Blue band	-100 - 16000	0.0001
Surface Reflectance band 4	Green band	-100 - 16000	0.0001
Surface Reflectance band 5	NIR band	-100 - 16000	0.0001
Surface Reflectance band 6	Short Wave Infrared (SWIR)	-100 - 16000	0.0001
Surface Reflectance band 7	SWIR	-100 - 16000	0.0001
Reflectance band quality	32-bit unsigned integer for describing quality of band pixels	NA	NA
Solar Zenith Angle	Sun zenith angle of band pixels	0 -18000	0.01
View Zenith Angle	View zenith angle of band pixels	0 -18000	0.01
Relative Azimuth Angle	Relative azimuth angle of band pixels	-18000-18000	0.01
State Flags	Quality of the product whether clouds or snow detected or not. (16-bit unsigned integer)	NA	NA
Day of Year	Julian day of the year	1- 366	NA

Table 2.3: Landsat-8 brief description

Band Number	Band Name	Spatial Resolution	Description
Band 1	Coastal/Aerosol	30 m	Coastal water and ocean colour analysis
Band 2	Blue	30 m	Discriminates vegetation from soil and deciduous from coniferous vegetation
Band 3	Green	30 m	Peak vegetation
Band 4	Red	30 m	Vegetation slopes. The vegetation absorbs the red band
Band 5	NIR	30 m	Focuses on biomass content and the band is reflected by the vegetation
Band 6	SWIR-1	30 m	Distinguishes between vegetation and soil moisture content
Band 7	SWIR-2	30 m	Improved difference between vegetation and soil moisture content
Band 8	Panchromatic	15 m	Better imaging capability
Band 9	Cirrus	30 m	Thin cloud detection
Band 10	TIR-1	100 m	Estimated soil moisture and thermal mapping
Band 11	TIR-2	100 m	Improved estimated soil moisture and thermal mapping

Sensor (TIRS) with a spatial resolution of 100 meters. Bands 1-9 (Coastal aerosol, Blue, Green, Red, NIR, SWIR 1, SWIR 2, Panchromatic, and Cirrus) are captured by OLI, and bands 10-11 (TIRS 1, TIRS 2) are captured by TIRS [39]. The description of each band is given in Table 2.3.

For Cloud cover assessment, Landsat 8 uses the CFMask algorithm for the identification of clouds, cloud shadow, snow, and ice and their representation in the QA band. Decision trees are used in CFMask for labeling the pixels in the scene. The cloud shadow mask is created by iteratively computing cloud height and projecting it onto the ground [40].

2.3.3 Sentinel-2

The spatial resolution for Sentinel-2 is 10m -60m and a revisit time of 10 days. After the launch of Sentinel-2B, the revisit time changed to 5 days with the same spatial resolution. The time series retrieved from the Sentinel-2 data can contain more useful patterns because of more number of repetitions in a year. Sentinel-2 covers the entire earth's land surface, islands, and coastal regions with the help of its two satellites [41]. The satellite provides 13 spectral bands and the description of these bands is given in Table 2.4.

The features of the area under consideration can be better analyzed by various combinations of the bands. The set of Band 2, 3, and 4 i.e. RGB (Natural color) represents the image as perceived by the human eye in which green color depicts vegetation, blue represents water, and grey or white shows the urban area. Bands 5-8a have a range of Visible and Nera Infrared (VNIR) and 9-12 are Shortwave Infrared (SWIR). The combination of bands 8, 4, and 3 (Colour Infrared) distinguishes between healthy and unhealthy vegetation in which red color depicts healthy vegetation due to the reflectance of NIR (band 8) by chlorophyll. Short-wave Infrared, a combination of bands 12, 8a, and 4 helps in the discrimination between vegetation (represented by green color) and bare soil (brown color). Bands 11, 8, and 2 combined help in monitoring crop health. Moisture Index can be computed using bands 8a and 11. The moist vegetation will have a high value for moisture index as compared to the dry vegetation.

Table 2.4: Band description of Sentinel-2

Band Number	Band Name	Spatial Resolution	Description
Band 1	Coastal Aerosol	60 m	For aerosol detection
Band 2	Blue	10 m	For soil and vegetation discrimination
Band 3	Green	10 m	Differentiates between clear and muddy water. Highlights oil on water surfaces, and vegetation
Band 4	Red	10 m	Identifies vegetation types, soils, and urban (city and town) areas
Band 5	VNIR	20 m	For classifying vegetation
Band 6	VNIR	20 m	For classifying vegetation
Band 7	VNIR	20 m	For classifying vegetation
Band 8	VNIR	10 m	Detecting and analysing vegetation
Band 8a	VNIR	20 m	For classifying vegetation
Band 9	SWIR	60 m	Detecting water vapour
Band 10	SWIR	60 m	Cirrus cloud detection
Band 11	SWIR	20 m	Measures moisture content of soil and vegetation
Band 12	SWIR	20 m	Differentiate between snow and clouds

2.4 Satellite Data Download and Pre-processing

Earth is divided into grids and each grid cell is called a tile. The satellite data is available in the form of tiles of varying sizes depending on the spatial resolution of the satellite. Each tile may contain data for multiple geographical locations e.g. counties/districts and vice-versa a location spread over multiple tiles. Collecting this data and mosaicking it for the required geographical location is a difficult and tedious task. It can also lead to redundancy of information or add irrelevant data if mosaicking is not done correctly. To solve this problem and make the data downloading process easier and faster, we have used Google Earth Engine (GEE) [37] which allows us to perform these tasks using its in-built functions and scripts. The data can be accessed from GEE in the form of 'ImageCollections'. A unique image collection ID is associated with each type of satellite e.g. there is a separate collection ID for MOD09V6, Landsat-8, and Sentinel-2. These image collections contain the data for the entire globe from the date of launch of the satellite. One has to filter out the data in terms of locations and time duration as per the requirement.

Some key steps are to be taken care of - filtering the required time duration, clipping the area of interest, removing the cloudy pixels, etc. To filter the data for the area of interest, administrative boundaries for the region are required to be uploaded on GEE. The data is obtained in the form of geotiff images.

2.4.1 Data Preprocessing

Satellite data often encounters challenges due to cloudy pixels and missing data, which can degrade the quality and reliability of analyses. Cloud cover obstructs the observation of Earth's surface, leading to incomplete or obscured information in satellite imagery. Additionally, missing data can arise due to sensor malfunctions, orbital constraints, or atmospheric conditions, further complicating the interpretation and utilization of satellite observations. Addressing these issues requires advanced data processing techniques such as cloud masking, interpolation, and data fusion to mitigate the impact of cloudy pixels and missing data on downstream applications.

We have handled both the problems as given below:

- *Handling cloudy pixels:* The scenes with less than 15 percentile of cloud cover are considered at every pixel using the simple composite algorithm of the Google Earth Engine.

- *Handling missing values:* The missing data at pixels is estimated using nearest neighbor interpolation and missing data for timestamps is estimated using linear interpolation. We used nearest-neighbor interpolation for pixel-level missing data. It involves replacing missing pixel values with the values of the nearest neighboring pixels because neighboring pixels often have similar values due to spatial autocorrelation. Also, this method is quick and computationally efficient. The missing data across timestamps is estimated using linear interpolation because it estimates missing values based on a linear relationship between known values at different timestamps [42].

2.5 Satellite images and derived Data

Satellite data has been used in various forms in different applications. It can be used as images of all bands, histograms, and spectral reflectance indices. The pictorial representation of the three forms is given in Figure 2.2. In addition to the pre-processing steps mentioned in section 2.3, there are some pre-processing steps specific to a particular way of modeling the satellite data. A detailed description of modeling ways and pre-processing is given in subsequent subsections.

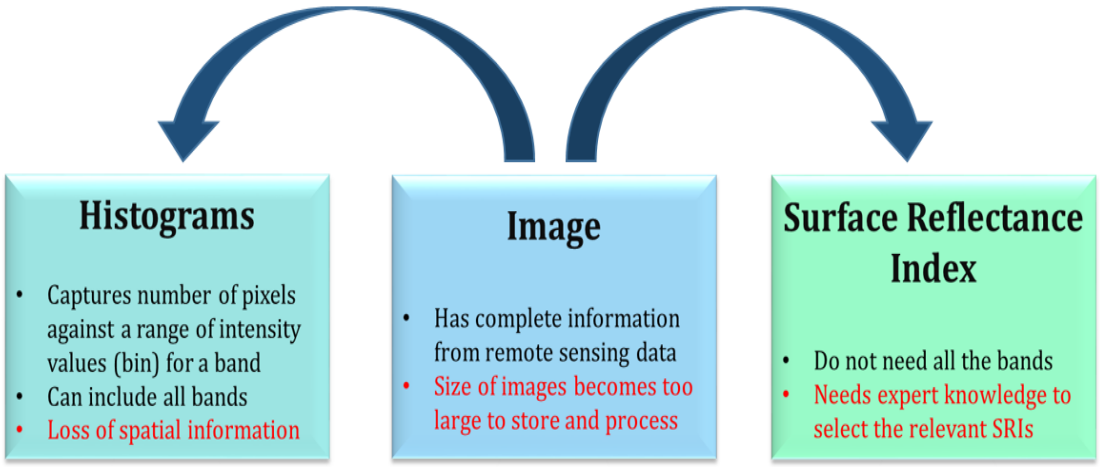

Figure 2.2: Different forms of satellite data

2.5.1 Images

The best to use maximum information from satellite data for an application is the use of the entire image as the images correctly capture spatial information. The size of a satellite image for a location is large depending upon the spatial and spectral resolution of the satellite system used.

Applications such as the prediction of crop yield, snow cover, soil moisture, etc. require time series analysis and analysis of historical data. Also, the ground truth for these problems is publicly available at the district or county levels. This requires the satellite

data to be captured and processed at the same level. For instance, in the prediction of crop yield, time series analysis is essential for capturing the seasonal variations, growth stages, and responses to environmental factors exhibited by crops over time. By analyzing historical satellite image time series, researchers can track the development of crops from planting to harvest, observe how environmental conditions such as temperature, precipitation, and soil moisture fluctuate throughout the growing season, and identify patterns that correlate with variations in yield. Similarly, in monitoring snow cover, time series analysis enables the detection of seasonal changes, trends, and anomalies in snow extent and depth over time. By analyzing historical satellite imagery, researchers can track the onset and duration of snowfall, observe patterns of snowmelt and accumulation, and assess the spatial distribution of snow cover across different regions. Thus time series analysis serves as a foundational tool for extracting actionable insights from historical satellite data across various applications.

Thus these applications require analysis of satellite image time series. The size of a single satellite image is large and combining these images captured at different time stamps further increases the volume of data to be processed. Thus, processing the image time series for such applications is challenging. To the best of our knowledge, there is very little work available in the literature in which raw satellite images have been used for earth observation applications over a large geographical area like a district or a county. For example, the number of pixels for a county in Landsat-8 is 2000×2000. Most researchers have worked at the pixel level and mainly for classification problems [25, 26] where ground truth is easily available. .

2.5.1.1 Data Pre-processing

The data preparation steps performed for the satellite image time series are given below:

- *Bits Precision:* Bits Precision is performed only for Landsat-8 images. By default, the Landsat-8 images have float values at every pixel for all the reflectance bands. Images in float values require 32 bits to store a single pixel. The number of pixels in a Landsat-8 image on average are 2000×2000. Thus it requires a large storage

space for an image to store. To work with the mentioned spatiotemporal applications, we need historical data for all the locations, thus storing so many images of such huge size is difficult. To save storage space we used the bits precision compression technique and converted all the float values to unsigned integer values as:

$$intvalue = round(floatvalue * 255) \tag{2.1}$$

It requires only 8 bits leading to a reduction in storage space by four. We verified the conversion by performing a set of preliminary experiments and the details are given in chapter 4 section 4.8.

2.5.2 Histograms

To overcome the computing bottleneck in working with satellite images, researchers have converted the images into histograms and used the histogram time series to solve problems like CYP, SCP, SEP [43, 44] etc. A histogram captures the information on the number of pixels against a range of intensity values (bin) for a band. For example, the count in bins is sufficient to predict the crop yield for a region. The particular locations of healthy/unhealthy crop plants in a region will not affect the aggregate yield of the region. According to the permutation in-variance assumption [43], it is the value of the pixel that contributes to the yield prediction for a specific location and not the position where that pixel is placed in the image. As shown in Figure 2.3, satellite images are converted into histogram volume of size $T \times B \times D$ where *B* is the number of bins, *T* is the number of timestamps and *D* is the number of bands e.g. obtained data volume of Landsat-8 having 23 timestamps for a year, for 64 bins and 9 bands is $23 \times 64 \times 9$. In this way 3D data is converted into 2D data for band. It is important to maintain the trade-off between the number of bins and the computation time required to process the time series. We have compared the impact of using different numbers of bins on crop yield prediction accuracy and the supporting results are shown in chapter 3, section 3.12.2.

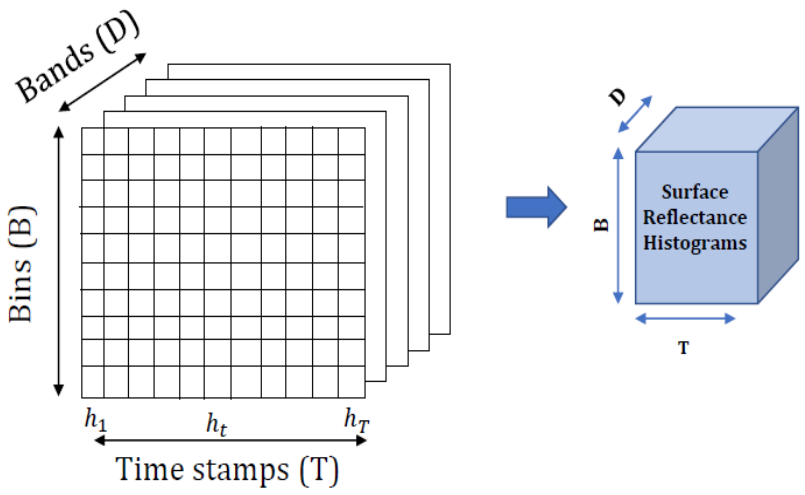

Figure 2.3: Histogram representation

2.5.2.1 Data Preprocessing and Histogram Creation

The size of multi-spectral satellite images is huge because of the high spatial resolution and the large number of bands. Moreover, we require historic time series data with weekly/fortnightly time granularity leading to a huge set of high-resolution images to be processed. Therefore, to handle computational cost (time and memory) we use data in the form of histograms for the spatiotemporal problems which subsume entries of images into intensity bins and convert an image at a timestamp for a band to a vector of size *B*. Satellite surface reflectance images are represented as histogram volume of size $T \times B \times D$ where *B* as shown in Figure 2.3. The soil data is also converted into histograms of dimension $1 \times B_s \times D_s$ having 1 timestamp as the soil is time-invariant for a location, B_s as the number of bins, and D_s as a number of depth levels. The pictorial representation of the complete process of data-preprocessing and histogram creation is shown in Figure 2.4.

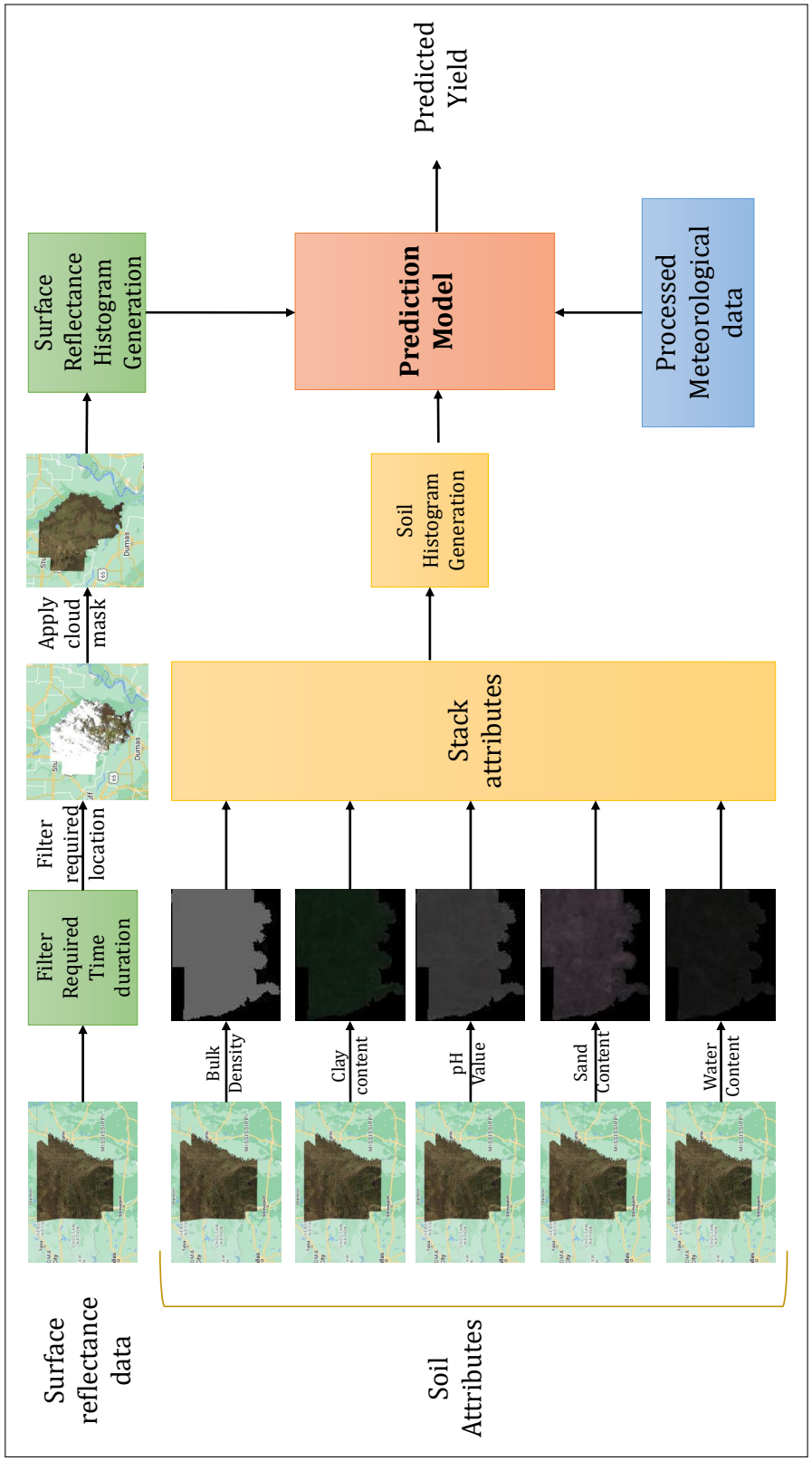

Figure 2.4: Histogram Creation

2.5.3 Spectral Reflectance Indices

Another method opted by researchers to handle the computational bottleneck in working with satellite images is to convert the satellite data into Spectral Reflectance Indices (SRIs). SRIs designed by physicists and domain experts, are mathematical formulations of bands in the visible and near-infrared electromagnetic spectrum. The formulations of these bands improve the sensitivity towards the detection of vegetation, environmental variables, physiological and morphological characteristics of the earth's surface or plants, etc. [45]. The use of SRIs has consistently proven to be a simple yet effective approach for assessing various aspects of vegetation, including qualitative and quantitative measures of growth parameters and health assessment. For example, Normalized Difference Vegetation Index (NDVI), derived from Red and Near Infra-red (NIR) bands has emerged as one of the most commonly employed spectral indices for monitoring crop growth and crop yield. However, various other SRIs have been used to support crop yield estimations like the Normalized Difference Water Index (NDWI), Enhanced Vegetation Index (EVI), etc. [46]. Different SRIs highlight different characteristics depending upon their mathematical formulations and thus can be useful in different applications.

We have derived all the spectral indices used from MODIS product MOD06 which has a spatial resolution of 500m and a revisit time of 8 days. The details of the SRIs considered are given below:

1. *Normalized Difference Vegetation Index (NDVI):* NDVI is an indicator of vegetation health based on how plants reflect certain ranges of the electromagnetic spectrum. Low NDVI values indicate poor health of vegetation and higher values indicate a higher density of green vegetation. The formula for NDVI is:

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)} \tag{2.2}$$

2. *Normalized Difference Water Index (NDWI):* NDWI, is used to differentiate water from dry land or is rather most suitable for water body mapping. Water bodies have

low radiation and strong absorbability in the visible infrared wavelength range. NDWI is sensitive to changes in the liquid water content of vegetation canopies. It is complementary to, not a substitute for NDVI. Values lie between -1 to +1, depending on the surface water content. High values of NDWI correspond to high water content. Low NDWI values correspond to low water content.

$$NDWI = \frac{(Green - NIR)}{(Green + NIR)} \tag{2.3}$$

3. *Soil Adjusted Vegetation Index (SAVI):* SAVI is a spectral reflectance index that attempts to minimize soil brightness influences using a soil brightness correction factor. This is often used in arid regions where barren land is more. The values of NDVI and SAVI change in the same pattern. Negative values are either water or urban areas. The higher the NDVI values (the same stands for SAVI) the denser (and more healthy) the vegetation. But NDVI starts saturating after the value of 0.7, while SAVI at this point is only 0.3. This means that SAVI can be better used in dense vegetation because it saturates slower than NDVI.

$$SAVI = \frac{(NIR - Red)}{(NIR + Red + L)} * (1 + L), L\ is\ constant \tag{2.4}$$

4. *Normalized Difference Yellowness Index (NDYI):* NDYI computed from the green and blue wavebands and overcomes limitations of the NDVI. It provides information about the yellowness of the target surface, with higher values indicating a greater intensity of yellow coloration.

$$NDYI = \frac{(Green - Blue)}{(Green + Blue)} \tag{2.5}$$

5. *Plant Senescence Reflectance Index (PSRI):* This index maximizes the sensitivity of the index to the ratio of bulk carotenoids (for example, alpha-carotene and beta-carotene) to chlorophyll. An increase in PSRI indicates increased canopy stress

(carotenoid pigment), the onset of canopy senescence, and plant fruit ripening. Applications include vegetation health monitoring, plant physiological stress detection, and crop production and yield analysis. The values of this index range from -1 to 1, with the common values for green vegetation ranging between -0.1 and 0.2

$$PSRI = \frac{(Red - Green)}{NIR} \tag{2.6}$$

6. *Enhanced Vegetation Index (EVI):* EVI is similar to NDVI and can be used to quantify vegetation greenness. However, EVI corrects for some atmospheric conditions and canopy background noise and is more sensitive in areas with dense vegetation.

$$EVI = G \times \frac{(NIR - Red)}{(NIR + C1 \times Red - C2 \times Blue + L)} \quad where\ G,\ C1,\ C2,\ L\ are\ constants \tag{2.7}$$

7. *Simple Ratio (SR):* This index is a commonly used spectral index in remote sensing for assessing surface characteristics and the wavelength of the deepest chlorophyll absorption across various environments. The simple equation is easy to understand and is effective over a wide range of conditions. In areas with homogeneous surface characteristics, such as bare soil regions, the SR tends to be close to 1. As the properties of the surface change, such as the presence of different materials or land cover types, the SR value also changes accordingly. However, it's important to note that the SR index is not bounded, and its values can exceed 1.

$$SR = \frac{NIR}{Red} \tag{2.8}$$

8. *Wide Dynamic Range Vegetation Index (WDRVI):* This index is similar to NDVI, but it uses a weighting coefficient (a) to reduce the disparity between the contributions of the near-infrared and red signals to the NDVI. The WDRVI is particularly

effective in scenes that have moderate-to-high vegetation density when NDVI exceeds 0.6. NDVI tends to level off when vegetation fraction and leaf area index (LAI) increase, whereas the WDRVI is more sensitive to a wider range of vegetation fractions. WDRVI enables a more robust characterization of crop physiological and phenological characteristics. Although this index needs further evaluation, the linear relationship with vegetation fraction and much higher sensitivity to change in LAI will be especially valuable for precision agriculture and monitoring vegetation status under conditions of moderate-to-high density. It is anticipated that the new index will complement the NDVI

$$WDRVI = \frac{(a \times NIR - Red)}{(a \times NIR + Red)} \ , 0.1 \leqslant a \leqslant 0.2 \tag{2.9}$$

9. *Modified Soil-Adjusted Vegetation Index (MSAVI):* The modified soil-adjusted vegetation index (MSAVI) is an index designed to substitute NDVI and NDRE where they fail to provide accurate data due to low vegetation or a lack of chlorophyll in the plants. It addresses some of the limitation of NDVI when applied to areas with a high degree of exposed soil surface. During the stages of germination and leaf development, there is a lot of bare soil between the seedlings. NDVI and NDRE both interpret this as poor vegetation. Here is where MSAVI comes to aid. "SA" stands for "soil-adjusted," revealing the key aspect of this vegetation index. It reduces the effect of the soil on the calculation of vegetation density in the field MSAVI values range from -1 to 1, where

   - -1 to 0.2 indicate bare soil
   - 0.2 to 0.4 is the seed germination stage
   - 0.4 to 0.6 is the leaf development stage
   - When the values go over 0.6, it is now high time to apply NDVI instead. In other words, the vegetation is dense enough to cover the soil

$$MSAVI = \frac{2 * NIR + 1 - \sqrt{(2 * NIR + 1)^2 - (8 * (NIR - Red))}}{2} \tag{2.10}$$

10. *Modified Soil Ratio (MSR):* MSR produces images with a good contrast. It was also observed that the MSR image has a better signal-to-noise ratio than that of the NDVI image. This could aid in making a reliable mapping of the vegetation cover of the area under study. MSR is proposed for retrieving biophysical parameters of boreal forests using remote sensing data. This SRI is formulated based on an evaluation of several two-band vegetation indices, including the Normalized Difference Vegetation Index (NDVI) and simple Ratio (SR).

$$MSR = \frac{\frac{NIR}{Red} - 1}{\sqrt{\frac{NIR}{Red}} + 1} \tag{2.11}$$

2.5.3.1 Creating Spectral Reflectance Index Image

The raw images of MODIS have 5 bands - Red, Green, Blue, NIR, and SWIR. After all the required corrections for missing data, we calculated the value of each spectral reflectance index at every pixel, thus forming a matrix for each vegetation index at a timestamp. Then we stacked all the vegetation index matrices of a timestamp one after the other making an image with 10 channels. The process is shown in Figure 2.5.

2.6 Meteorological Data

We used different meteorological attributes including maximum temperature, minimum temperature, Radiance, precipitation, Dew, Humidity, Visibility, Wind speed, Cloud cover, sea level pressure, wind gust, solar radiation, and solar energy.

2.6.1 Data Preprocessing

Meteorological data also requires pre-processing steps listed below.

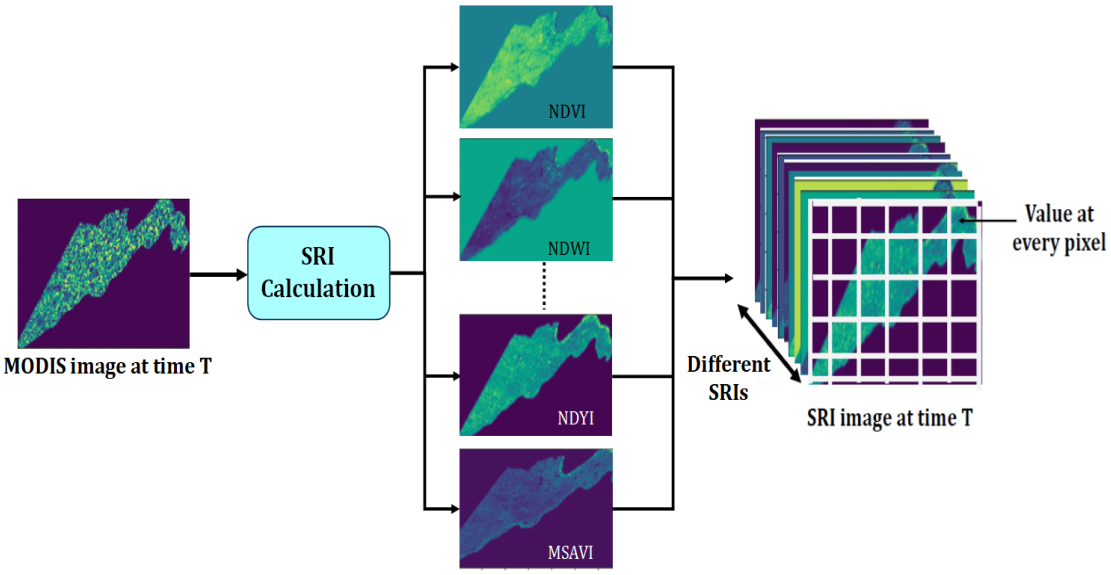

Figure 2.5: SRI image at a particular timestamp T

2.6.1.1 Handling Missing Values

The meteorological data is recorded on a daily basis and it also suffers from missing data. The method of forward fill, a common approach for handling missing values in time series data is used to estimate missing values at any day. This assumes a certain continuity in the time series, where the most recent known value is deemed a reasonable estimate for the missing value.

2.6.1.2 Temporal granularity

Since the meteorological data is captured at a higher temporal granularity than satellite data, an aggregation step is necessary. The meteorological data is averaged to align with the temporal resolution of the satellite data, ensuring compatibility between the two datasets. This aggregated and imputed meteorological data can then be integrated with the satellite data, allowing for cohesive analysis and model training.

2.7 Soil Data

The soil properties do not change with time for a location and thus the soil data does not have any temporal resolution and can be downloaded for the required spatial resolution. We downloaded soil data from Open Land Map [47] for different soil properties. The soil

properties for which the data is available include – clay content, bulk density, pH in water, water content, carbon content, and sand content. The soil data is captured at 6 different depth levels from ground level– 0, 10, 30, 60, 100, and 200 (in cm). So, this makes a total of 6 levels for each soil attribute for a location. The data can be used in the form of images or histograms. We used it in the form of histograms as shown in Figure 2.4.

Chapter 3

Crop Yield Prediction: An Important Earth Observation Application

[1]

3.1 Introduction

Agriculture is one of the most important industrial sectors and the backbone of economic development and food security in any nation. Crop yield prediction (CYP) is the most crucial for agriculture. Yield prediction refers to the estimate of the crop produced in a year based on historical data. It plays a vital role in decision-making by the government and farmers. Early and timely yield prediction helps the government in making better decisions for import/export, warehouse management for post-harvest of crops, etc. [48].

Predicting crop yield accurately is a challenging task as it is dependent on various meteorological factors like climate (maximum and minimum temperature, rainfall, precipita-

[1] *The work presented in this chapter has resulted in the following publications:*

- *Arshveer Kaur, Poonam Goyal, Kartik Sharma, Lakshay Sharma, and Navneet Goyal, "A generalized multimodal deep learning model for early crop yield prediction", in IEEE International Conference on Big Data 2022.*
- *Arshveer Kaur, Poonam Goyal, Rohit Rajhans, Lakshya Agarwal, and Navneet Goyal, "Fusion of multivariate time series meteorological and static soil data for multistage crop yield prediction using multi-head self attention network", in Expert Systems with Applications (2023).*

tion, humidity, etc.), soil (soil type, groundwater availability, soil moisture, soil ph-value, etc.), location, seed variety and resources available to farmers. The factors affecting the yield vary both temporally and spatially, thus making the CYP problem, a spatiotemporal problem. However, researchers have attempted to model crop yield prediction in many ways. Authors in many studies have taken meteorological data as static data along with soil and genotype data. Out of weather, soil, and genotype parameters, weather parameters are the ones having the maximum variability during a crop cycle. Suitable weather conditions at every stage of the crop are important for its growth and better yield. Several machine-learning models have been used for CYP. The commonly used models are random forest (RF) [49], support vector regression (SVR) [50], multi-layer perceptron (MLP) [51], k-nearest neighbor (KNN), support vector machine (SVM) [52], etc. Deep learning models used in the literature include deep neural networks (DNN) [52–57], convolutional neural network (CNN) [43, 44, 48, 58], RNN [44, 48], LSTM [43, 58], etc.

In recent times, remote sensing data has gained the attention of researchers, due to its easy availability and capability to scale up across the regions. There are many satellites viz., MODIS [10], Landsat [40], Sentinel [41], etc. providing remote sensing data which is available in different temporal, spatial, and spectral resolutions. Every satellite gives data for a fixed number of reflectance bands.

The existing studies for the CYP problem have used MODIS satellite data which has a spatial resolution of 500m and temporal resolution of 8 days. Other satellites, Landsat-8 and Sentinel-2, launched in the later years (2013 and 2015, respectively), have higher spatial resolution than that of MODIS. The data from Landsat-8 and Sentinel-2 satellites is available only for 7 and 5 years, respectively (2021 and 2022 yield data for crops not available) and thus has not been yet used for CYP.

In this chapter, we have worked with three modalities of data viz. meteorological, soil, and surface reflectance bands. In the first part, we have taken meteorological and soil data collected in a conventional (not remote sensing) way for crop yield prediction. In the second part, we have used conventionally collected meteorological data along with satellite-based soil data and surface reflectance bands. We conducted experiments to

establish the fact that adding satellite modality to earth observation applications improves the model accuracy. From the next chapter onwards we have focused primarily on satellite data for various earth observation applications.

3.2 CYP using conventionally collected data

Many researchers have worked on solving the problem of CYP in recent times. The existing studies have modeled the problem in different ways using statistical models, machine learning, and deep learning models using numeric data which includes meteorological data, genotype, soil, and remote sensing data. Researchers have used these data individually or in combination to solve the CYP problem.

3.2.1 Related Work

The simplest way to estimate yield is using only the meteorological data. Verma et al. [59] have used climate variables including rainfall, maximum and minimum temperature, humidity, etc. for predicting the yield of the mustard crop in the Haryana state of India. The authors used the principle component analysis and multiple linear regression to predict yield at the district level. Authors in [60] have also used the meteorological data for wheat yield prediction with the World Food Studies (WOFOST) model.

The authors [61] applied a simple regression model for predicting the rice yield based on only weather data. A Seasonal Prediction System (SPS) is used to predict the weather attributes using historical data. This predicted weather data is used in the Crop System Model CERES-Rice to predict the rice yield. The accurate prediction of yield depends on the accuracy of the predicted weather data. The errors and anomalies in the weather prediction model are carried forward to the yield prediction model leading to a poor estimate of yield for those years.

Ensemble model of Ada Boost with SVM and Naïve Bayes classifiers is used to predict the yield of different crops including Sugarcane, Rice, Cotton, Groundnut, etc. using only the climatic data [62]. The weather parameters considered include maximum and minimum temperature, precipitation, vapor pressure, cloud cover, and wet day frequency,

captured at the temporal granularity of the month.

Droesch [53] applied a semi-parametric deep neural network for predicting the corn yield in the US Midwest area from the years 1976-2016. The data used included the past yield data for corn and historic weather parameters (maximum and minimum temperature, wind speed, precipitation, humidity, and radiation) collected daily. The model performed better than the statistical and simple machine learning models as it was able to capture the non-linear relationship between the weather attributes and yield.

Sharma et al. [63] proposed an artificial neural network using a Bayesian optimization approach for corn yield prediction based on meteorological factors like temperature, precipitation, and geographic coordinates of the location.

The authors [64] have applied ML models viz. random forest, neural network, and SVM for estimating the rice yield. The experiments were carried out on minimum, maximum, and mean values using different weather and soil parameters including day-time temperature, night temperature, precipitation, humidity, wind speed, soil moisture, and soil temperature averaged at two different spatial granularity levels: district level and taluk level. The study shows that taking the data at a finer taluk level was more beneficial for accurate yield prediction.

The authors have modeled the yield prediction problem as a classification problem by classifying the yield into three classes' viz. low, mid, and high [52]. The yield is predicted using three classifier models - k-nearest neighbor, SVM, and least squared vector machine using past yield data, soil parameters, and rainfall in the area.

Khaki et al. [48] proposed a CNN-RNN model for yield prediction of corn and soybean using the weather and soil parameters. The weather parameters included maximum and minimum temperature, vapor pressure, precipitation, snow water equivalent, and solar radiation. The soil parameters include bulk density, pH value, organic matter percentage, hydraulic conductivity, water content, etc.

Other than using weather and soil data, genotype data also plays an important role in the yield of a crop. Few researchers have incorporated the genotype data as well

for CYP. Khaki et al. [65] in another work designed a deep neural network for maize yield prediction using meteorological data, soil, and genotype data. The data used in the paper is Syngenta crop challenge data (not available now) consisting of 8 and 6 soil and weather parameters, respectively along with 2267 genotype hybrids planted over different locations. The weather data is averaged over month granularity making in a total of 72 weather parameters.

The authors [66] have applied the LSTM model on the dataset consisting of weather parameters and genotype details for soybean yield prediction. The weather data is taken as a time series on a weekly basis. The LSTM model was reported to perform better than support vector regression with radial basis function kernel (SVR-RBF) and least absolute shrinkage and selection operator (LASSO). The study is done to find the most suitable genotype for a location.

Maloy et al. [67] proposed a deep learning model for yield prediction of barley using genotype and environmental data. The study was done to make the breeding decisions for different genotypes depending on the meteorological data and yield of crops using the specific genotype in particular weather conditions.

Research Gaps: All the existing studies lack incorporating the data from different modalities and appropriate modeling of the CYP problem. We hypothesize that locations with 'similar' meteorological and soil conditions will have 'similar' yield patterns. The existing models for CYP are mostly trained with the meteorological data of the entire year. However, most of the crops do not have their crop cycles spanned over the whole year, and thereby training models with data that is not directly relevant leads to poor training. The existing models predict the yield using the meteorological data for the entire crop cycle and prediction is made at harvest time. Predicting the crop yield at the harvest time is not as useful as predicting the yield during the crop cycle, many weeks before the harvest time.

3.2.2 Study Area and Data Used

We used NC94 numeric data which consists of 4 meteorological and 102 soil attributes. It is collected at the county level covering 1055 counties from 12 states highlighted in the map (Figure 3.1). We used some additional numeric meteorological attributes. The description of the dataset is given in Table 2.1. We have taken three crops viz. soybean, wheat, and corn for experiments. The number of counties taken for each crop is 603, 345, and 628 respectively. The counties are selected based on a crop cultivated there and the availability of the weather data for the entire crop cycle. Some of the selected counties can be common across crops. The meteorological and soil attributes have a varying range of values and units in NC94. We normalize all the attributes using min-max normalization to remove any kind of biases.

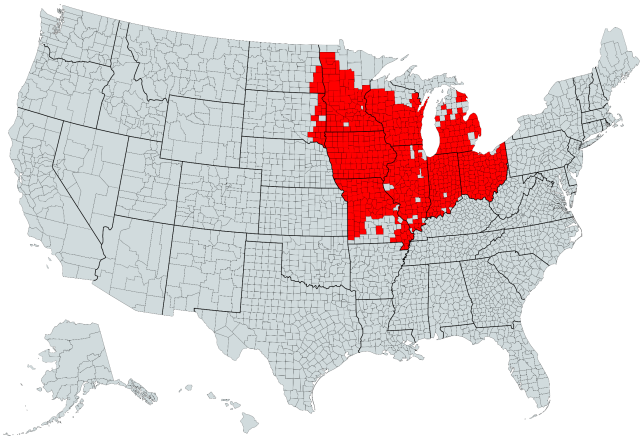

Figure 3.1: Study Area: Numeric Data (YieldPredictNet)

3.2.3 Modelling of Problem: Numeric Data

The pipeline of the system used for yield prediction is given in Figure 3.2. The raw data is first pre-processed for handling missing values and aggregating them as per the required temporal granularity of week or month. The next module performs the spatial clustering of locations (counties) w.r.t. meteorological and soil characteristics. The backbone of the pipeline is a sequential model realized by LSTM units with multi-head self-attention for time series data for learning temporal relationships. We use two variable selection

modules (i) Attribute Selection Unit and (ii) Depth-level Selection Unit (DLSU) for static soil attributes. In the end, we predict the crop yield.

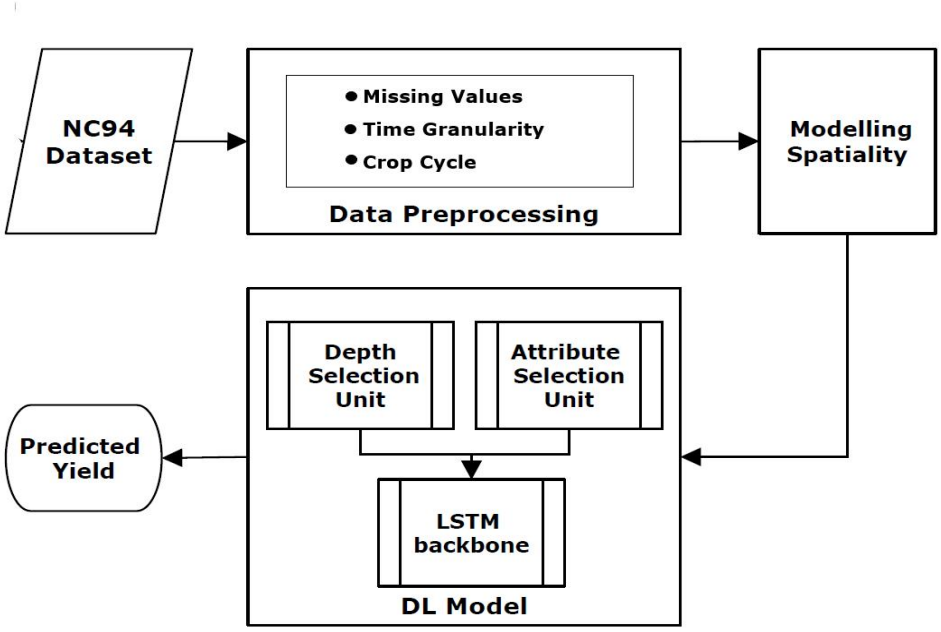

Figure 3.2: Modelling the crop yield prediction problem

3.2.3.1 Padded Crop Cycle

The meteorological data has a significant impact on the growth and yield of a crop. Most existing studies are done using the weather data for the entire year [48, 55, 61, 65, 68–70]. The crops experience different crop cycle starting at different times of the year. In the US, the crop cycle for soybean starts in mid-May and lasts till November end. The crop cycle of corn lasts from mid-March to mid-November and for wheat sowing starts in mid-September and harvesting takes place at the end of July. Few studies have also used the crop cycle [43, 44, 56, 58] rather than the entire year data. We have considered three lengths of time series- (i) entire year (Y), (ii) weeks or months covering the crop cycle (CC), and (iii) padded crop cycle (PCC). In the padded crop cycle, we have padded two extra weeks on either end of the crop cycle. Padding the crop cycle helps in modeling the discrepancy in the sowing and harvesting time at different locations. Padding also helps in capturing any anomaly or sudden change in climatic conditions before the sowing of

a crop. In real-time, the weather conditions before the sowing of the crop can have a substantial impact on the growth of the crop as it can affect the soil conditions as well.

3.2.4 Modelling Spatiality

Yield prediction is a spatiotemporal problem because meteorological and soil data vary with locations. In addition, Meteorological data has very fine time granularity, but soil data can be considered largely static. We have considered meteorological data at weekly granularity. We have exploited the spatiality of the problem by clustering the counties. We hypothesize that the counties experiencing 'similar' weather and soil conditions tend to have 'similar' yield patterns. NC94 dataset consists of the geographical locations spread over a large area of North Central US. There is a massive disparity in the soil characteristics and meteorological conditions experienced by different states and counties in this area. The counties are clustered using the k-means algorithm (i) using only meteorological data and (ii) using soil and meteorological data. We have fine-tuned the proposed base deep learning model by further training it with the data of individual clusters. In this way, the model can learn both global and cluster-specific yield patterns. The results prove that our hypothesis is correct as the error in predicted yield reduces significantly after training the model on cluster-specific data. We have also clustered the data using Dynamic Time Warping (DTW) distance measure and achieved comparable results. The clustering is done for all three lengths of the time series i.e. Y, CC, and PCC. The obtained k-means clusters for weekly PCC using soil and meteorological data (giving the best results) for different crops are presented in Figure 3.3. The number of clusters obtained for soybean, wheat, and corn are 3, 4, and 3, respectively shown in different colors. The number of clusters taken in each crop is decided with the help of the inverse scree method [71]. We have not considered too many clusters as it would reduce the data for each cluster and consequently, the base deep learning models would not get fine-tuned properly.

3.2.4.1 Modelling Temporality

The data is a multivariate time series as all the meteorological attributes vary with time. The length of the time series depends on the granularity level and its span depends on

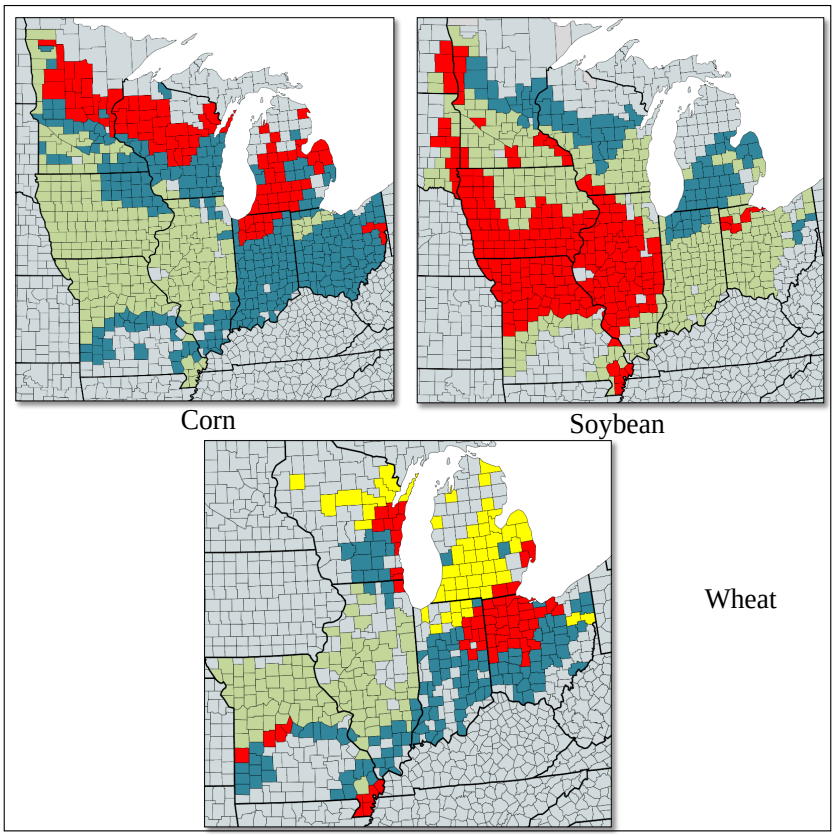

Figure 3.3: K-means clusters obtained for weekly PCC using meteorological and soil data

whether we are considering data for the entire year, crop cycle, or padded crop cycle. The length of the time series for soybean, wheat, and corn is 25, 44, and 35 respectively for crop cycle data. The time series for meteorological data is represented as $M = \{M_1, M_2, ..M_t, ...M_T\}$ where T is a number of time stamps, $M_t = \{m_{it}\}, i = 1$ to q is the set of q meteorological attributes at any time *t*.

3.2.4.2 Attribute Selection and Depth Selection for Soil Variables

In the NC94 dataset, the soil data was collected only once as the soil properties change very marginally. Therefore, soil attributes can be considered largely static in comparison to meteorological data. The dataset contains 102 soil attributes out of which 11 attributes are depth-variant and is collected at 6 different depths and 36 attributes are depth-invariant. We designed an Attribute Selection Unit (ASU) to select a subset of *k* attributes that play an important role for CYP from a total of *n* input parameters. We have performed attribute selection in two ways: (i) Flat Attribute Selection (FAS) - Out of all

the $n = 102$ attributes top k relevant attributes are selected. (ii) Soil Depth Modelling (SDM) - In this case, FAS is done only for $n = 36$ depth-invariant attributes. We have modified and used ASU to select the appropriate depth till at which the parameter is relevant for CYP for each depth-variant soil attribute at each time stamp (as shown in Figure 2.1 in Chapter 2). For example, i^{th} soil factor, S_i, is selected at depth 3 this means that the attributes s_{i0}, s_{i1} and s_{i2} will be considered for CYP at that time stamp. We have named this module as Depth-level Selection Unit (DLSU). This exercise may help farmers to focus on enhancing the relevant soil properties with the help of fertilizers or pesticides if possible. It is evident from the results that modeling the soil variant attributes has a significant impact on predicting the yield accurately.

3.2.5 Model Architecture: YieldPredictNet

The architecture of the proposed model YieldPredictNet is shown in Figure 3.4. The proposed model has three modules. The LSTM module consists of LSTM units with multi-head self-attention as the backbone of the system to process the time series. It learns the temporal dependencies present in the data. The other two modules are the Attribute Selection Unit (ASU) and Depth-level Selection Unit (DLSU) given in Figure 3.5 and described later in this section.

Model Input: We input meteorological data as multi-variate time series and soil data as static. The model inputs the meteorological data directly into the LSTM module. The soil attributes are input to either ASU or both ASU and DLSU for FAS and SDM, respectively.

3.2.5.1 LSTM-module

The multivariate time series meteorological data is modeled as a sequential model using LSTM units. LSTM is a type of RNN, with forget gate [72]. It helps LSTM to decide when and what information is to be forgotten. The gates and the cell states in LSTM make it suitable for dealing with the long-term dependencies and problem of vanishing gradient. The output of the module is the predicted crop yield for the year (N+1) based on its learning from the past N years. We use bidirectional LSTM that enables learning

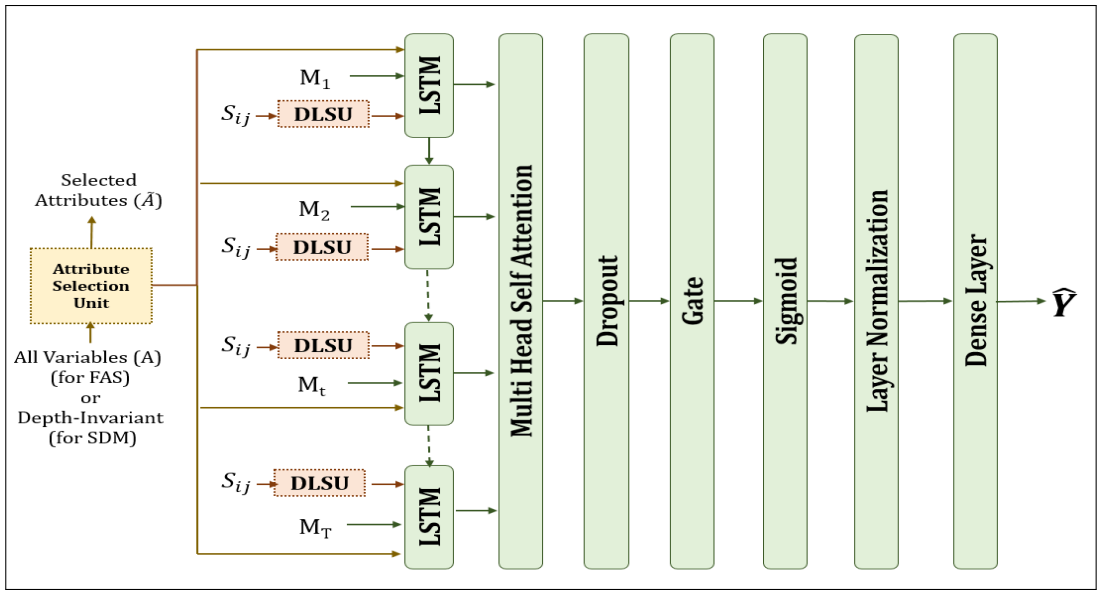

Figure 3.4: The Model Architecture: YieldPredictNet

in both directions, forward and backward. In the backward direction, learning is done from the future to the past. The combination effectively learns the relationships involving attributes and yield at any timestamp during the crop cycle and is thus helpful in predicting yield at the early stages of the crop cycle. The output from LSTM units is passed to the Multi-head Self Attention module. The futile layers in the network are skipped using the gate mechanism. The sigmoid activation function is used after the gated layer and then layer normalization is applied to normalize the feature for having a mean of zero and variance as one. Layer normalization helps in faster convergence of sequential models. As the layer normalization works on an instance basis, it avoids any kind of dependencies between the batches making it suitable for sequential models.

Multi-head Self Attention: Attention is used to exploit and learn the long-term dependencies in the time series or sequential input. The attention layer takes three parameters as input viz., Query (Q), Key (K), and Value (V) in the form of vectors. It is referred to as a mapping of query, key-value pairs to the output [73]. The attention is calculated with the help of a query by comparing it with the keys to get weights for the values. Self-attention means having the same value for key, query, and value i.e. K=Q=V. The self-attention can be used in three ways – self-attention in an encoder, self-attention in a

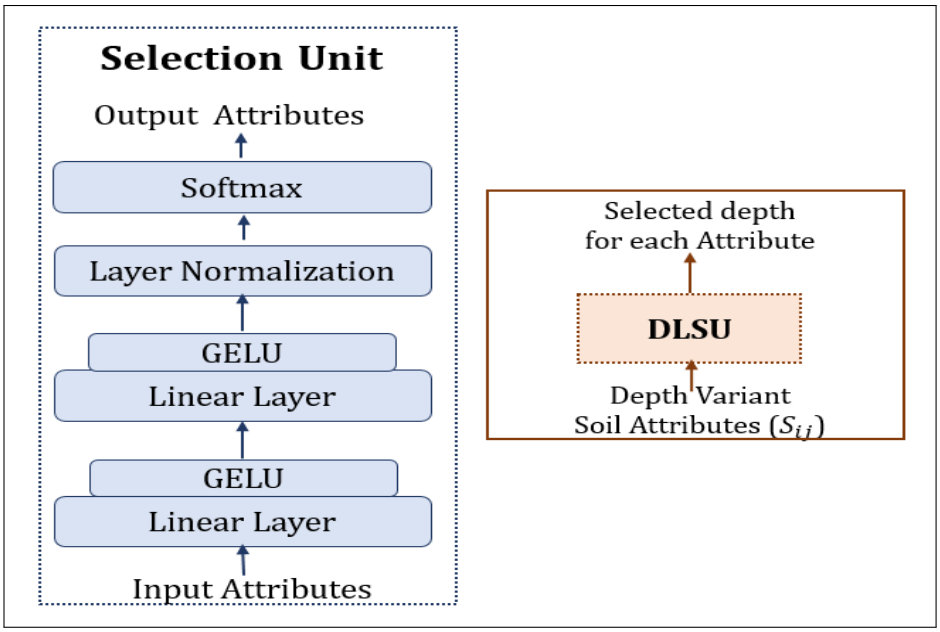

Figure 3.5: Attribute Selection Units

decoder, and encoder-decoder attention in the decoder. We have used self-attention in the encoder as CYP is a sequence-to-number problem and the decoder is not involved.

Multi-head attention refers to the mechanism of dividing the input to be computed in parallel on different multiple heads. The heads output the different representations for each (query, key, value) pair. The output from each head is joined to find the final attention score or weight assigned to every element in the time series. The computational cost is equal to that of a single head because of the reduced dimensions at each head. The mathematical representation of attention is given below:

$$Attention(Q, K, V) = Softmax((QK^T)/\sqrt{d})V \tag{3.1}$$

where *d* denotes the dimension of the Key and is used as a scaling factor for preventing the dot product from having larger values in magnitude. The input is divided into *m* parts to be executed at *m* different heads denoted by *h*. The attention at each head is denoted as below:

$$h_i = Attention(Qw_i^Q, Kw_i^K, Vw_i^V) \tag{3.2}$$

where w_i^Q, w_i^K and w_i^V represent the weights assigned by i^{th} head to query, key and value,

respectively. The output from every head is concatenated as:

$$Multihead(Q, K, V) = [h_1, h_2, \ldots., h_m]W^o \tag{3.3}$$

3.2.5.2 Attribute Selection Unit (ASU)

This module is intended to select the depth-invariant soil attributes relevant to CYP. The input to ASU is the set of *n* soil attributes denoted by $A = \{a_1, a_2, \ldots.a_n\}$ out of which the unit selects *k* most relevant attributes. ASU consists of two linear layers having a Gaussian Error Linear Unit (GELU) activation function, a Layer Normalization layer, and a softmax layer to output weights for each attribute. LayerNorm is a standard normalization layer for normalizing the activation results of the previous layer for every instance of the data. Layer normalization helps in removing the dependencies among the instances. GELU is an activation function [74] which combines the properties of zone out, dropout, and Rectified Linear Unit (ReLU) for intensifying the probability of neuron output. GELU shows the curvature at every point because of its non-monotonic and non-convex property. The mathematical formulation of the module is given below.

$$\tau_1 = W_1 A + b_1 \tag{3.4}$$

$$\tilde{\tau_1} = GELU(\tau_1)k \tag{3.5}$$

$$GELU = xP(X \leqslant x) = x\phi(x) \tag{3.6}$$

$$= x/2(1 + \tanh[(2/\Pi)(x + 0.044715x^3)] \tag{3.7}$$

where ϕ is the cumulative distribution function for Gaussian distribution.

$$\tau_2 = W_2(\tilde{\tau_1}) + b_2) \tag{3.8}$$

where (W_1, b_1) and (W_2, b_2) are the weights and bias at two linear layers, respectively.

$$\tilde{\tau_2} = GELU(\tau_2) \tag{3.9}$$

$$\delta = LayerNorm(\tilde{\tau_2}) \tag{3.10}$$

$$\gamma = Softmax(\delta) \tag{3.11}$$

$$= \sum_{i=1}^{n} \gamma_i \times a_i \tag{3.12}$$

where γ_i represents the softmax weight associated with each attribute a_i, τ_1 and τ_2 are the two linear layers, and $\tilde{\tau_1}$ and $\tilde{\tau_2}$ depicts the intermediate form of linear layers τ_1 and τ_2 after applying GELU. The output of the unit will be $\tilde{A}$, the set of selected *k* attributes.

3.2.5.3 Depth Level Selection Unit (DLSU)

The module selects the appropriate depth level for each soil factor since soil property is relevant to crop growth up to a certain depth. The significance of the depth level for a factor may vary with different stages of the crop cycle. That is why, we use a separate DLSU at each time stamp of the crop cycle (i.e. phenological stage of a crop). The input S_i where, $i = 1$ to p and $S_i = s_{ij}, j = 1$ to z to the DLSU is *p* soil factors captured at z different depths. The output of the unit is selected depth level l_i for each factor S_i. l_i is the depth-level up to which the attribute S_i is suitable for a crop at a time stamp. DLSU has a similar composition as that of ASU. It consists of two linear layers with a GELU activation function, followed by a Layer Normalization layer and a softmax layer similar to ASU. The final softmax layer will be modified as:

$$\gamma = \sum_{i=1}^{p} \sum_{j=1}^{z} \gamma_{ij} \times s_{ij} \tag{3.13}$$

where γ_{ij} represents the weight given to each attribute s_{ij} The final output of the DLSU unit will be the selected depth-level l_i for each attribute $s_i (i = 1 \text{ to } p)$ at each phenological stage of crops.

$$\tilde{S_i} = \{s_{1l_1}, ..., s_{pl_p}\} \tag{3.14}$$

The selected attributes from DLSU and ASU are passed to the LSTM module.

3.2.5.4 Forecasting

The prediction is done in two ways: *end of season* prediction and *in-season* prediction. In most of the studies, prediction is done at the end of the year or crop cycle. A single value is given as the output by the linear layer applied at the end of all the functions and transformations. The output $\hat{Y}$ represents the predicted yield of the crop. The forecasting model can be represented as:

$$\hat{Y}_{N+1} = F([\mathbf{M_1}, \mathbf{M_2}, ..., \mathbf{M_N}], \tilde{A}, \tilde{S}_i) \tag{3.15}$$

where $\hat{Y}_{N+1}$ is the yield predicted by the model for $(N+1)^{th}$ year and F denotes the forecasting model and M_i is set of meteorological attributes at N^{th} year.

Early Prediction: For early prediction, we have modified our end-of-cycle model to predict yield at the end of any week of the crop cycle. The error is back-propagated at every timestamp rather than only at the last time stamp. The early prediction is done at 4 different *in-season* time stamps. The yield is predicted at intervals of 2 weeks starting from 8 weeks before the harvesting.

3.2.6 Experimental Setup for YieldPredictNet

The implementation is done using Pytorch open-source library in Python. 4 Tesla M10 GPU servers with 8 GB of VRAM are used to run the experiments. Hyper-parameter optimization is done using the Adam optimizer for all three crops viz. soybean, corn, and wheat. The clustering of counties is done to exploit the spatiality patterns. The counties are clustered into 3, 4, and 3 clusters for soybean, wheat, and corn, respectively. The hidden dimensions and learning rate are 80 and 0.0000045, respectively when only meteorological data is used and 1000 and 0.0000025, respectively when both soil and meteorological data are used. We have taken a batch size of 64 for all experiments. The number of epochs for experiments using only meteorological data and using both meteorological and soil data are 100 and 50, respectively. When the data is taken for the entire year, the length of the time series will be 52 for all the crops. If only the crop cycle is considered, the length is 25, 44, and 35 for soybean, wheat, and corn, respectively. If a padded crop cycle is considered, then the crop cycle length increases by four time stamps for all crops (2 on either end of the crop cycle).

3.2.7 Evaluation Metric

We have used the mean squared error as the loss function for the prediction model. The evaluation metric used is root mean square error (RMSE). The formula for RMSE is given

below in Equation 7 where y_i depicts the actual yield, $\hat{y_i}$ denotes the predicted yield for a county in a year, and *C* denotes the number of counties considered for each crop. The yield is measured in bushels/acre (bu/ac).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{C}(y - \hat{y_i})^2}{C}} \tag{3.16}$$

Training and Testing: The prediction is done in two ways by taking single-time horizons and multi-time horizons. In taking a single time horizon, prediction is done for $(N+1)^{th}$ year using the data of preceding *N* years. While predicting yield for $(N+2)^{th}$ year, the models are trained for previous *N*+1 years and so on. In the case of the multi-time horizon, we predict the yield for years (*N*+1) to $(N+x)$ using the model trained only for the first *N* years. We have considered $x = 5$ in our experiments.

3.2.8 Models for comparison: YieldPredictNet

We have done a comparison of YieldPredictNet with six existing models for yield prediction which use non-remote sensing data. All the models are optimized and trained on the NC94 dataset for every crop separately in the same way as we have done for YPN for fair comparisons. The models taken for comparison are:

Random Forest [39]: The number of base estimators for random forest taken in the case of soybean, wheat, and corn are 11, 9, and 13, respectively and the maximum depth for each decision tree of random forest is 11, 17, and 9, respectively.

LASSO [66]: Least Absolute Shrinkage and Selection Operator (LASSO) is a type of regression model with a regularization technique. The shrinkage factor alpha (taken as 0.1) helps in shrinking towards the center point of the data.

Support Vector Regression (SVR) [66]: SVR works similarly to SVM but the margin is approximated in the case of SVR because the output for regression problems is a real number and has infinite possibilities. The kernel function used is the radial basis function.

DNN [58]: DNN model predicts the yield by taking both meteorological and soil data as static data after flattening it. The number of layers and neurons used at every layer is (11,25), (11,19), and (14,33) in the case of soybean, wheat, and corn, respectively. The model uses the Adam optimizer for hyperparameter tuning with a learning rate of 0.0001.

CNN-RNN model [48]: The model takes both the meteorological and soil data as the multivariate time series data. The model uses a W-CNN with 60 neurons for learning weather data features an S-CNN with 40 neurons for soil data features and finally an LSTM layer with 64 cells for predicting the yield for the target year. It uses stochastic gradient descent as the optimizer with a learning rate of 0.001.

LSTM [66]: The model uses meteorological (time series) data and genotype (static) data for prediction. We have used the soil data in place of the genotype as both are static and don't require any modification to the model. The model uses an Adam optimizer with a learning rate of 0.001.

3.2.9 Experimental Scenarios for YieldPredictNet

All the experiments are executed for six scenarios and are denoted by acronyms given in Table 3.1. The first scenario M considers only the meteorological data for both training and prediction. While both meteorological and soil data are considered in all the other scenarios from 2-6. The model taking entire meteorological and soil data is referred to as MS. For model CM, clustering is done only on meteorological data. CM+RMS represents clustering on meteorological data and then retraining the MS model for each cluster to exploit cluster-specific patterns. Similarly, model CMS denotes that the model uses both meteorological and soil data for clustering and CMS+RMS denotes retrained MS using CMS clusters.

3.2.10 Results: YieldPredictNet

NC94 dataset consists of 30 years of data, out of which we have taken 5 years of data (1996-2000) for testing. The results are captured at two different time granularities (weekly and monthly) and three different lengths of time series viz. Y, CC, and PCC.

Table 3.1: Different scenarios considered for experiments

Sce. No.	Met. Data	Soil Data	Met. Clusters	(M+S) Clusters	Retrain Model	Model
1	✓	×	×	×	×	M
2	✓	✓	×	×	×	MS
3	✓	✓	✓	×	×	CM
4	✓	✓	✓	×	MS	CM+RMS
5	✓	✓	×	✓	×	CMS
6	✓	✓	×	✓	MS	CMS+RMS

Table 3.2: Average RMSE of the proposed model (YPN) for all the crops

Soybean								
Week Granularity								
TS length	M	MS	CM	CM+RMS	CMS	CMS +RMS	Add. Data CMS+RMS	DTW CMS+RMS
Y	8.3042	5.6733	5.7868	5.7238	5.7859	5.634	5.3897	5.5111
CC	8.5514	5.7014	5.9277	5.6514	5.8495	5.5195	5.3228	5.4077
PCC	7.6866	5.64	5.6137	5.4204	5.5578	5.4565	**5.3172**	5.3842
Month Granularity								
Y	14.6052	9.5231	8.3252	8.1542	9.1255	7.5236	5.6878	5.7582
CC	12.1795	8.1957	6.8524	6.3712	6.5242	6.2112	5.4921	5.5207
Wheat								
Week Granularity								
TS length	M	MS	CM	CM+RMS	CMS	CMS +RMS	Add. Data CMS+RMS	DTW CMS+RMS
Y	14.6901	11.2957	11.5223	11.1144	11.4686	10.9496	10.7971	10.5106
CC	13.8013	11.1694	12.1415	11.8496	12.2059	10.8134	9.907	10.2245
PCC	14.2361	10.9011	10.8802	10.8052	10.79	10.6162	**9.8814**	10.1913
Month Granularity								
Y	15.7983	14.646	15.6833	11.6682	15.6191	11.6187	11.0073	11.1729
CC	15.0086	12.3899	14.1266	11.3703	14.1024	11.3124	10.7971	10.8691
Corn								
Week Granularity								
TS length	M	MS	CM	CM+RMS	CMS	CMS +RMS	Add. Data CMS+RMS	DTW CMS+RMS
Y	29.039	23.0627	22.7516	21.4038	22.2438	20.9824	18.6348	19.2052
CC	29.4	21.3086	22.4019	21.7268	22.2723	19.9613	17.9576	18.7616
PCC	27.1383	20.4639	20.7716	19.8635	20.9189	19.7816	**17.4857**	18.429
Month Granularity								
Y	45.2208	36.3924	33.3699	28.3718	33.6849	28.4192	24.7902	24.9587
CC	32.6111	30.8778	30.626	25.2128	30.6282	25.1494	22.8676	22.9226

Ablation study: We have done extensive experiments considering a number of design choices of the proposed model to better understand their relative importance. Table 3.2 presents the impact of using the type of data and the kind of modeling used for the problem for different six successive models (see Table 3.1 for description). The results show that taking soil along with meteorological data, substantially improves the accuracy as compared to the results obtained with only meteorological data for all lengths of time series and crops. Further, adding additional meteorological attributes reduces the RMSE up to 11% across the crops as they play a role in the growth of the crop. It can also be observed from Table 3.2 that RMSE is the least when a padded crop cycle is used. This is observed for all crops and other data choices considered in this chapter.

The counties are clustered for modeling spatiality and the clusters are obtained using *k*-means with L_2 and DTW distance. The experiments show that the average RMSE obtained with DTW distance is slightly higher in comparison to that of clusters obtained using L_2 distance across the crops (see Table 3.2). Clustering reduces the data for training the cluster-specific models and leads to poor training. Re-training the MS model on cluster-specific data leads to a significant reduction in the error. This allows the model to learn both global and cluster-specific local patterns.

The re-trained CMS+RMS model with additional meteorological attributes and L_2 *k*-means clusters on both climate and soil data surpass the performance of all the other scenarios for all crops. The results also show that the granularity at which the data is collected also plays an important role in accurately predicting the crop yield. Taking the data at week granularity predicts the yield more accurately by approx 3%, 8.5%, and 23% for soybean, wheat, and corn, respectively than taking the data at month granularity.

We have also tested our design choices on existing methods (Table 3.3). We observed that all the observations for design choices hold true for all the existing models. The given results establish that considering soil attributes has a positive impact on yield prediction. Based on the above observations, we present all results of our proposed model with additional meteorological data at week granularity with PCC length of time series and *k*-means clustering with L_2 distance. We call this design choice 'YieldPredictNet'.

Table 3.3: Validating the design choices with existing models (RMSE)

	Soybean			Wheat			Corn		
	M	MS	CMS+RMS	M	MS	CMS+RMS	M	MS	CMS+RMS
RF Y [39]	9.9581	8.3624		16.8628	16.7173		32.0968	28.8766	
RF PCC [39]	9.6276	8.2058		16.6636	16.359		32.0785	28.7598	
LASSO Y [66]	9.8989	9.4993		15.9859	14.9161		32.6857	32.1811	
LASSO PCC [66]	9.8144	9.2491		15.8545	14.8458		31.2857	31.2283	
SVR Y [66]	8.969	8.9113		15.4256	15.1953		29.6903	29.8375	
SVR PCC [66]	8.9111	8.3674		15.2042	15.0482		29.3855	29.8364	
DNN Y [65]	10.2308	9.9314	9.5253	15.9499	15.6917	15.5337	31.8609	30.9158	29.7353
DNN PCC [65]	9.7342	8.3486	7.9004	15.7192	15.5311	15.2334	31.7515	30.4815	28.6908
CNN-LSTM Y [48]	8.5438	7.5215	6.3667	15.186	14.9869	13.5027	30.2309	28.4722	26.4279
CNN-LSTM PCC [48]	8.1481	7.4182	6.0599	14.9988	14.4203	11.8602	29.1197	25.6651	24.3317
LSTM Y [66]	11.7993	10.7993	9.1442	18.1381	17.4734	15.1088	34.9841	32.4939	29.9097
LSTM PCC [66]	11.6602	9.7665	8.9299	17.6821	15.9637	15.0528	33.2012	32.4004	29.6173
YPN Y	8.3042	6.7652	5.6898	14.6901	11.2957	10.7971	29.039	23.0627	18.6348
YPN PCC	8.0408	5.9014	**5.3172**	14.2361	10.9011	**9.8814**	27.1383	20.4639	**17.4857**

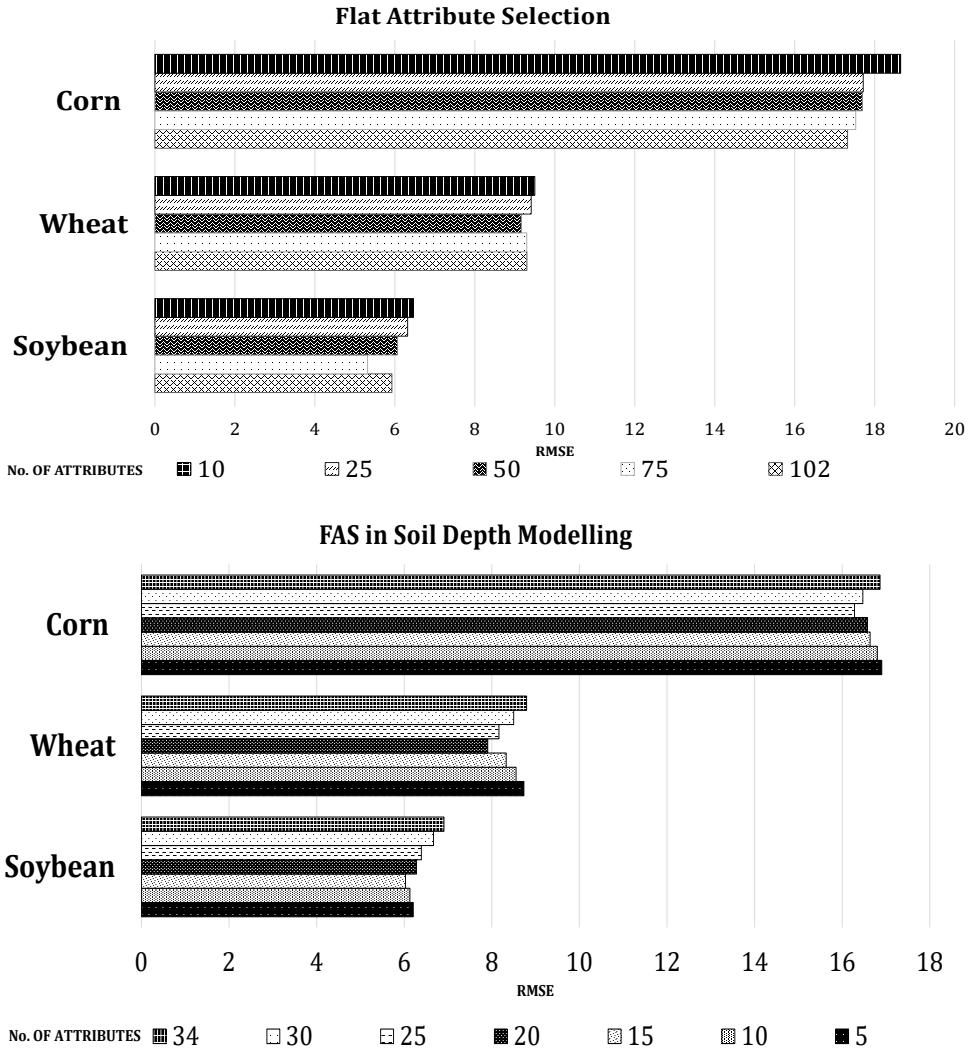

Figure 3.6: Variation in RMSE with number of soil parameters in the proposed model (YPN) with FAS and SDM

As per attribute selection of soil attributes as described in section 3.2.4.2, we have compared the two attribute selection units viz. FAS and SDM in Figure 3.6. Table 3.4 shows the difference in results while taking all the attributes versus FAS and SDM. We observed the variance in the performance of the model by varying the number of soil attributes selected for FAS and SDM (Figure 3.6). It can be observed that the number of favorable attributes for different crops are different e.g. in FAS the least RMSE is achieved by taking 75, 50, and 102 attributes for soybean, wheat, and corn, respectively.

All the experimental results presented are taken by considering the optimal number of soil attributes for all the crops in every scenario. FAS gives the improvement of up to 7% improvement in yield prediction over using all the features (Table 3.4).

Table 3.4: Comparison of YPN with existing models on NC94 dataset (RMSE)

	Soybean	Wheat	Corn
DNN [65]	7.9004	15.2334	28.6908
CNN-LSTM [48]	6.0599	11.8602	24.3317
LSTM [66]	8.9299	15.0528	29.6173
YPN-AF	5.3172	9.8814	17.3193
YPN-FAS	5.3050	9.1613	17.3193
YPN-SDM	**5.2551**	**7.8654**	**15.5408**

The results given in Figure 3.7 show that the relevant depth of different factors is different throughout the crop cycle. The graph shows the depth level after every 15 days. The required depth of a factor differs with respect to crop as well. E.g., the depth required for measuring the percentage of organic matter present in the soil is around 50cm, 10cm, and 25cm, respectively for the soybean, wheat, and corn at starting of their respective crop cycles. Likewise, the appropriate level changes throughout the crop cycle. Modeling the depth-variant factors has shown significant improvement over using all attributes for yield prediction. The percentage improvement is approximately 1.17%, 20.40%, and 11.12%, for soybean, wheat, and corn, respectively.

We compare the performance of our proposed model YPN with six existing models given in Table 3.4. YPN- with AF, FAS, and SDM represent YPN with all the attributes taken, YPN with flat attribute selection, and YPN with soil depth modeling, respectively. It is evident from the results that our proposed model (YPN) predicts the yield more accurately as compared to other models in all three ways of modeling the soil attributes. The best results are achieved when depth-variant soil factors are modeled separately and a suitable depth level is selected for each factor at every time stamp.

Early Prediction: So far we have given *end of season* (end of PCC) prediction. We have modified our YPN model for early prediction. An early and accurate prediction

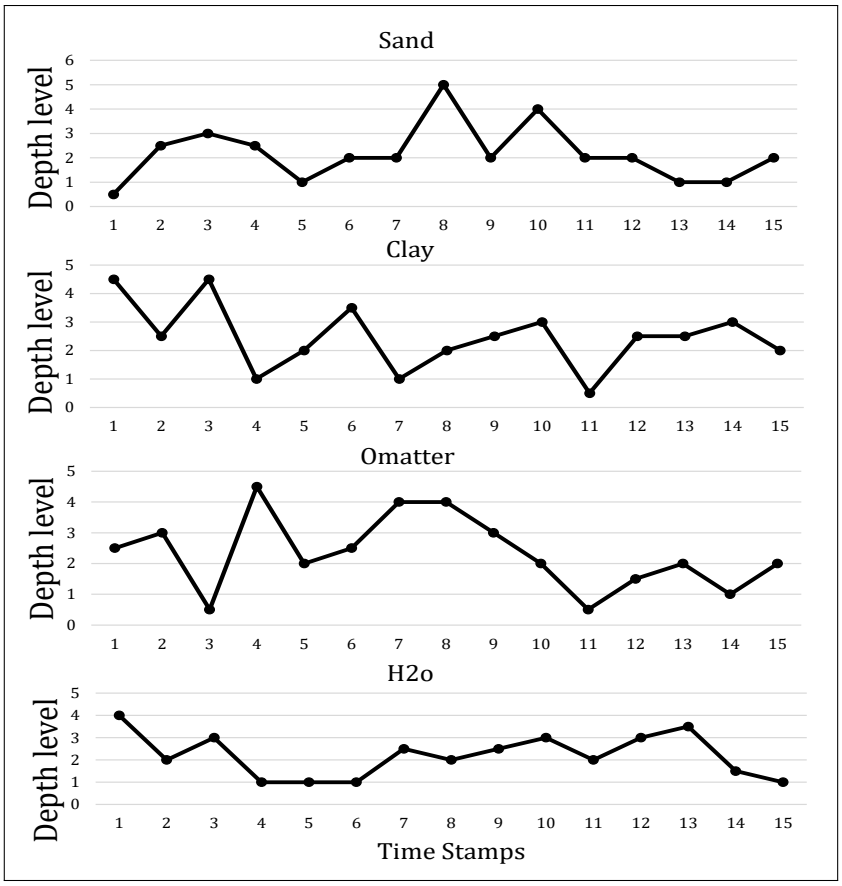

Figure 3.7: Soil depth required throughout the crop cycle for soybean. The corresponding depth levels are- 0:10cm, 1:25cm, 2:50cm, 3:100cm, 4:200cm and 5:250cm

can help farmers and the government to intervene appropriately if required. We have predicted the yield every two weeks backward from the harvest time (end of PCC) of the crop. The results for early prediction given in Table 3.5 show that the model can predict the yield with almost the same accuracy at 8 weeks before the harvest. As expected, the prediction accuracy is closer to harvest time. We have compared our results for early prediction with the available model in literature [58]. Our results at *T*-8 have an error of 0.26% from the prediction at time *T*. This is much lower than the 3% error of the existing model [58].

Multi-time horizon Prediction: In multi-time horizon prediction, we have trained the model till *N* years and predicted the yield for $(N+1)^{th}$ year to $(N+x=5)^{th}$ year. Table 3.6 shows the difference between single-time horizon prediction and multi-time horizon predictions. As expected, the percentage error increases from $(N+2)^{th}$ year to

Table 3.5: Early yield prediction using YPN-FAS and SDM

Prediction		Early Prediction				End of Season
Crop	Time Step	T-8	T-6	T-4	T-2	T
Soybean	FAS	5.8532	5.8514	5.8509	5.8498	5.8469
	SDM	5.2692	5.2666	5.2657	5.2626	5.2551
Wheat	FAS	9.1651	9.1624	9.1616	9.1614	9.1613
	SDM	7.8809	7.8795	7.8758	7.8726	7.8654
Corn	FAS	17.6009	17.5709	17.5577	17.5383	17.3193
	SDM	15.6315	15.5990	15.5841	15.5637	15.5408

Table 3.6: Multi-horizon Prediction
SHP: Single-time horizon prediction & MHP: Multi-time horizon prediction

	Soybean		Wheat		Corn	
Year	SHP	MHP	SHP	MHP	SHP	MHP
N+1	4.3716	4.3716	6.1242	6.1242	16.3457	16.3457
N+2	4.3075	4.5254	7.2327	7.3952	13.0571	13.9175
N+3	5.6265	6.0201	6.8254	7.4253	16.5955	17.8760
N+4	6.4556	6.9707	9.5241	12.7661	18.3569	19.8391
N+5	5.5147	6.1518	9.6204	13.5387	13.3486	15.1363

$(N+5)^{th}$ year. This can be attributed to the fact that meteorological conditions from $(N+1)^{th}$ year to $(N+(k-1))^{th}$ year are not factored in while predicting for $(N+k)^{th}$ year. This leads to the maximum error of approximately 10%, 29%, and 12% for soybean, wheat, and corn, respectively for $(N+k)^{th}$ year.

3.3 Crop Yield Prediction: Satellite Data

With the advancements in remote sensing data, it has been lately used in various agricultural tasks. The satellite data is captured in the form of multi-spectral images consisting of a different number of surface reflectance bands depending on the satellite.

3.3.1 Related Work

As mentioned in Chapter 2, the surface reflectance bands are either used as an image, histograms [43, 58], or some mathematical formula is applied over two or more bands to derive a SRI. The SRIs used in yield prediction include NDVI, enhanced vegetation index (EVI), Normalized difference water index, (NDWI), etc. [39, 44, 55, 69, 75]. The authors [39] used NDVI and NDWI derived from Sentinel data for wheat yield prediction in Madhya Pradesh (India). The random forest regression model achieved better accuracy as compared to other machine learning algorithms including SVM regression and multivariate polynomial regression. The authors [76] used the SRI along with meteorological data for wheat yield prediction with multiple models including random forest, Lasso, K-Nearest Neighbour, and Multi-layer perceptron (MLP).

As compared to the machine learning models, deep learning models perform better with satellite data. A few researchers have applied deep learning models in their work for CYP [44, 48, 53–55, 58, 65]. Although the satellite data is itself spatiotemporal, in the existing studies it has been either taken as only time series [44, 48, 55, 58] or by flattening it as static data [43, 48, 54]. Different deep-learning models have been applied for CYP using different kinds of data for various crops. CNN and RNN model [48] are used by taking weather and soil parameters as time series, a model designed by combining CNN and LSTM [44] for yield prediction by taking surface reflectance bands as time series and soil as static data.

Another deep learning Spike neural network (SNN) [55] is applied over the NDVI data to analyze crop health and phenological characteristics for yield prediction. The authors in [44] and [43] have tried to exploit the spatiotemporality of the satellite data but the SNN [55] model is incapable of handling the static data as it works on the concept of analyzing the spikes in the dataset and static data will be constant and not have any spikes.

Considering all the aspects and different types of data involved, we can say that CYP is a spatiotemporal problem, but in the existing studies it has been either modeled as only

time series [55, 58, 77] or static [43, 64, 78, 79]. Sun et.al. [44] have done CYP taking surface reflectance data as a time series and soil as static. The soil data used in the study is captured at varying depth levels for each attribute and this information is not dealt with.

Research Gaps: The existing studies working with satellite data have been focusing on remote sensing data captured by only MODIS satellites. The data from recent satellites with high spatial resolution is not considered for the application because of its availability for limited years. We hypothesize that high-resolution satellite data can provide better insights about the factors affecting crop yield, even when limited historical data is available for training suitably designed deep learning models. Moreover, the above-mentioned studies work only for a specific region or a specific crop.

3.3.2 Study Area and Data Used: Satellite Data

We used histograms created from three satellites viz. MODIS, Landsat-8, and Sentinel-2. The histogram creation process is explained in section 2.5.1. We have made yield predictions for the years 2019 and 2020. we used MODIS data from the year 2001, Landsat-8, and Sentinel-2 from 2014 and 2016, respectively. The data from Landsat-8 and Sentinel-2 satellites is available only for 7 and 5 years, respectively as they were launched in the later years (2013 and 2015, respectively).

Yield data: The yield data for crops is taken at the county level for the US from Quick Stats [80] collected by the United States Department of Agriculture (USDA) and at the district level for India [81]. We have taken data from year 2014 to 2020.

Study Area: The area considered for CYN includes the counties of the US and districts of India which are listed as the top producers of corn and soybean. The US counties include from states viz. Illinois, Indiana, Iowa, Kansas, Kentucky, Michigan, Minnesota, Mississippi, etc. The district data is taken from states of India viz. Andhra Pradesh, Telangana, Punjab, Rajasthan, Uttarakhand, Himachal Pradesh, Assam, etc.

3.3.3 Data Preparation

Handling missing values: The data could be missing for a few timestamps because of cloudy days, technical problems with the sensor of the satellite, etc. The missing data at certain time stamps is estimated using linear interpolation [42].

Processing meteorological data: The missing values in meteorological data are estimated using the forward fill method in which the missing value at timestamp T for an attribute is replaced by the value at timestamp T-1 for that attribute. The meteorological data is captured at daily granularity and is aggregated to match the temporal granularity of the respective satellite data as given in Chapter 2.

3.3.4 Model Architecture: CropYieldNet

We proposed CropYieldNet to capture information from data of different modalities such as numeric time series (meteorological), and static (depth-variant soil data) spatiotemporal (surface reflectance bands data) obtained from satellite systems. The proposed model, CYN, consists of four modules- Surface Reflectance Encoder (SRE) which learns the spatial patterns in surface reflectance data without tampering the temporal patterns, Soil Data Encoder (SDE) to learn the pixel intensity information across bins for each soil attribute, Depth-level Selection Module (DSM) to select the soil data input only till the relevant depth-level for each attribute, and Core Temporal Module (CTM) to exploit the temporal patterns in surface reflectance and meteorological data and ultimately predict the output yield. The pictorial representation of CYN architecture is given in Figure 3.8.

Note that all four modules are jointly learned in an end-to-end process via stochastic gradient descent and contrastive learning applied to satellite data.

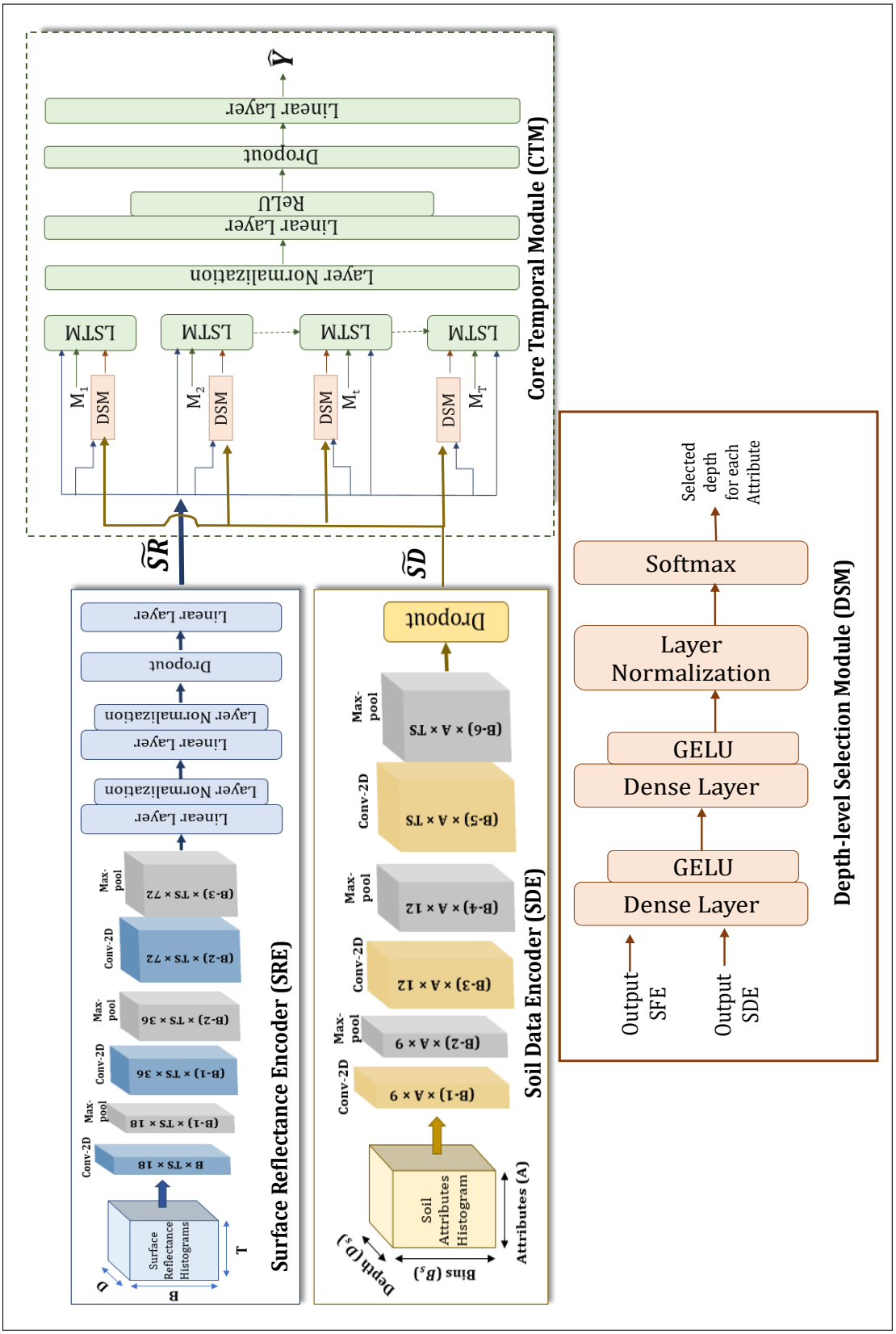

Figure 3.8: Proposed Model Architecture (CropYieldNet)

Model Input: An input instance to SRE is the data for a year/crop-cycle for a location. The surface reflectance data for a timestamp is stored as a histogram (see Histogram Generation of chapter 2). The histograms for all *T* timestamps are arranged in columns to form a 2D matrix of dimensions $B \times T$ for each band. The band information has been taken as channels and the data tensor, *SR*, for instance, has size $B \times T \times D$ (see Figure 2.3). Similarly, meteorological data for all timestamps in a year/crop cycle is arranged as a 2D matrix of size $n \times T$ and passed directly to the CTM module. The soil data is also converted into histograms of dimension $1 \times B_s \times D_s$ having 1 timestamp because soil is time-invariant for a location, B_s as the number of bins, and D_s as the number of depth-levels, and is passed to SDE.

3.3.4.1 Surface Reflectance Encoder (SRE)

This module is intended to learn the spatial patterns of the surface reflectance data. Let the data input be denoted by *SR*. The SR encoder consists of 1D CNN to capture spatial patterns in the data. It contains convolution layers with 18, 36, and 72, filters in successive layers. Each convolution layer is followed by a maxpool layer. To keep the temporal information intact, we have taken the kernel size as 2×1 with stride '1'. The last maxpool layer is followed by two blocks of linear, normalization, and ReLU activation layers. A dropout layer with a probability of 0.3 is used to regularize the model and prevent it from over-fitting. At the end, the linear layer is used to flatten the tensor of each timestamp resulting in 2D output of SRE which is denoted by $\widetilde{SR}$.

3.3.4.2 Soil Data Encoder (SDE)

Soil Data Encoder encodes soil information using CNN as the backbone. Soil information in the form of rearranged soil histograms is taken as a 3D tensor with shape $B_s \times 1 \times D_s$, where B_s is the number of bins and D_s as the number of depth levels for each soil attribute. Collective input for all attributes is taken as $B_s \times A \times D_s$, where *A* is the number of soil attributes. SDE consists of three 1D convolution layers with successive number of channels as 9, 12, and *T*. Each convolution layer is followed by a maxpool layer. The kernel size is taken as 2×1 to process the soil data for each attribute separately. The last maxpool layer is followed by a dropout layer with a probability of 0.3 to nullify the

contribution of some neurons toward the output. This is done to prevent the model from over-fitting and to prevent the biased influence of the first input batch in model training. The output of SDE is a 3D tensor denoted as: $\widetilde{SD}$ with dimensions ${B_s}' \times A \times {D_s}'$.

3.3.4.3 Depth-level Selection Module (DSM)

The soil attributes are collected at 6 different depth levels and it is possible that an attribute is not relevant at all depths at a timestamp for crop growth. We modified DLSU (section 3.5.3) to work with satellite-obtained soil data and named it as Depth-level Selection Module (DSM). DSM selects the appropriate depth level up to which a soil attribute contributes to crop growth. Depth level is selected based on the surface reflectance data representing the phenological stages of a crop. DSM takes two inputs- $\widetilde{SR}$ and $\widetilde{SD}$. We have rearranged the $\widetilde{SD}$ to make its dimensions identical to that of $\widetilde{SR}$.

DSM consists of two dense layers each followed by a Gaussian Error Linear Unit (GELU) [74] activation layer. GELU is considered a smoother ReLU because it combines the benefits of ReLU activation and dropout regularization. At the time of thresholding, it weighs the inputs by their value, unlike ReLU which weighs on the basis of their sign. GELU is followed by a normalization layer. Layer normalization is used to stabilize the hidden state dynamics in LSTM and make the learning process faster. Softmax is applied at the end to assign weights to each depth level and select the most relevant depth level (*d*) at every timestamp (*t*). A separate DSM is used for every timestamp. The output of DSM for a soil attribute is the data from level 1 to *d* of the attribute. This data corresponding to all A soil attributes is denoted by $\widetilde{SD'}$.

3.3.4.4 Core Temporal Module (CTM)

The Core Temporal Module is designed taking LSTM as its backbone followed by a linear layer with ReLU activation function and a dropout layer with probability 0.1. A linear layer is used at the end to output the predicted yield. LSTM is a recurrent network with forget gate which helps the model in deciding what information is to be forgotten at each timestamp. The cell states and forget gate in LSTM make it convenient to handle the long-term dependencies in the time series and deal with the problem of vanishing gradient. We have used a bidirectional LSTM model as it learns the patterns in data in both

forward and backward directions. This makes it suitable to learn the temporal patterns correctly from the augmented data generated using the inversion technique. Also, using the bidirectional LSTM helped in predicting precise in-season yield. The input to the module is output from SRE and DSM. The preprocessed meteorological data is directly input to the module. The meteorological data is represented as $M = \{M_1, M_2, ..., M_T\}$ where *M* and $M_t = \{m_{it}\}, i = 1$ to *n* are the set of meteorological data for *T* timestamps and at a timestamp *t*, respectively. m_{it} is set of meteorological attributes. The prediction model is denoted by:

$$\widehat{Y}_{N+1} = CTM(\widetilde{SR}, M, \widetilde{SD'}) \tag{3.17}$$

where $\widehat{Y}_{N+1}$ is the predicted yield by CTM for year $N + 1$ using the static soil data and *N* years of surface reflectance and meteorological data.

In-season Prediction: We modified our end-of-season model to predict in-season yield at multiple time stamps. We have back-propagated the loss at every time stamp of the model and thus trained the model at all the in-season stages of the crop cycle. We have done the early prediction from *T*-5 to *T* timestamps.

3.3.5 Data Augmentation: CropYieldNet

Deep learning models work remarkably well on time series data as compared to traditional machine learning models. However, these models require a large amount of data for training to get superior performance. The data from Landsat-8 and Sentinel-2 is available only for 7 and 5 years, respectively and crop yield data is also not available for most of the locations over the globe. Therefore data augmentation (DA) is essential for the success of the model. Unlike vision and NLP, DA in time series is more challenging to find the techniques that do not tamper with the intrinsic properties of the time series data and generate valid data which enhances the generalization capability of the model. We propose to apply two DA techniques to the surface reflectance band data of satellites.

Inverting Time Series: We have reversed the order of time series as (*T*, *T*-1,...,*t*,..,2,1)

for surface reflectance band and meteorological data and labeled with the same target yield. This will not affect adversely as our model learns time series through bidirectional LSTMs.

Temporal Irregularity: We generate training samples by dropping 10% of the timestamped data with 0.1 probability. keeping the same target value. This results in a differently shaped input data volume which is handled by padding the volume with zeros to match with the size of the input data.

3.3.6 Training objectives: CropYieldNet

We have used standard mean square error MSE and contrastive loss as training objectives for regression and contrastive learning, respectively.

Contrastive Loss: The revival of studies in contrastive learning has made major advancements in self-supervised representation learning [82]. We applied batch-wise N-pair contrastive loss to learn the mutual spatiotemporal patterns in different representations of the same surface reflectance and soil data instance. Two different representations of $\widetilde{SR}$ and $\widetilde{SD}$ are created by varying the dropout probability with 0.1 in SRE and SDE, respectively. It is done to create the positive pair for the instance (anchor). Corresponding negative pairs are selected from a set of other samples in the mini-batch. The working principle of contrastive learning is to pull the positive pairs closer to the anchor and push negative pairs away from it. The mathematical formulation of the process is given below.

$$\widetilde{SR}_{i1} = SRE(SR_i, dp_1) \tag{3.18}$$

$$\widetilde{SR}_{i2} = SRE(SR_i, dp_2) \tag{3.19}$$

where SR_i is the set of surface reflectance data for a county *i* for *T*, dp_1 and dp_2 are the dropout probabilities ($dp_1 \neq dp_2$). We will calculate the dot product between all the samples in one batch of size *k*. Let's say $\widetilde{SR}_{i1}$ as the original representation and $\widetilde{SR}_{i2}$ as augmented, *i* ranges from 1 to *k*. For each, $\widetilde{SR}_{i1}$, there is only one positive pair in $\widetilde{SR}_{i2}$ and all other 2k-1 are negative pairs.

$$pos = \widetilde{SR}_{i1}.dot(\widetilde{SR}_{j2}), \quad say\ i = j \tag{3.20}$$

$$neg = otherwise \tag{3.21}$$

Calculating the contrastive loss using the following equation:

$$CL_1 = -\log[\frac{exp(pos/\tau)}{\sum_{i=1}^{k} exp(\frac{neg}{\tau})}], \quad i \neq k \tag{3.22}$$

where *pos* and *neg* is the similarity between positive and negative pairs, respectively for surface reflectance data, τ denotes the temperature constant which helps in maintaining the gradient for the optimization process. We have taken $\tau = 27.17$ after hyper-parameter tuning [83].

Similarly, contrastive loss is calculated for soil data:

$$CL_2 = -\log[\frac{exp(pos_s/\tau)}{\sum_{i=1}^{k} exp(\frac{neg_s}{\tau})}], \quad i \neq k \tag{3.23}$$

where pos_s and neg_s are the similarities between positive and negative pairs, respectively for soil data.

Mean squared error (MSE): We have used the standard mean squared error (MSE) function as the loss function for regression. The formula for MSE is given below where y_i depicts the actual yield, $\hat{y}_i$ denotes the predicted yield for the county for the year, and *N* denotes the number of counties considered for each crop.

$$MSE = \sum_{i=1}^{N} \frac{(y - \hat{y}_i)^2}{N} \tag{3.24}$$

The total loss *L* for the model is:

$$L = CL_1 + CL_2 + MSE \tag{3.25}$$

3.3.7 Experimental setup for CropYieldNet

The implementation is done using Pytorch open source library in Python. A30 GPU server with 24 GB of VRAM is used to run the experiments. Hyperparameter tuning is done using Stochastic Gradient Descent with a learning rate of 0.0001 and momentum of 0.7. The hidden dimensions for LSTM in our baseline model are taken as 30. The sequence length of LSTM changes with the temporal resolution of the satellite and duration (T) considered for time series. If the time series is considered for the entire year (Y)/padded crop cycle (PCC), then T takes values 46/34, 23/15, and 37/25 for MODIS, Landsat-8 and Sentinel-2, respectively. We have used 5 and 3 years of data for training for Landsat-8 and Sentinel-2, respectively, and 2 (2019 and 2020) years of data for testing. To predict the yield for N^{th} year, training is done till $(N-1)^{th}$ year. For example, to predict for 2019, training is done from 2014-2018 and 2016-2018 for Landsat-8 and Sentinel-2, respectively. To predict yield for 2020, training is done till 2019.

3.3.8 Models for comparison: CropYieldNet

We compare our model (CYN) with three existing models CNN [43], CNN+GP [43], and CNN+LSTM [44], the first two of which uses only surface reflectance data and the third uses soil data with surface reflectance data. These models are trained and tested for different crops, locations, and time duration. However, we have used the data for the same locations and time duration in all the models for a fair comparison. The experiments for CNN and CNN+GP are performed only on surface reflectance data as the model is incapable of taking any other data as input. Similarly, CNN+LSTM can only take surface reflectance and soil data as input. The models CNN, CNN+GP, and CNN+LSTM use Adam optimizer for hyper-parameter tuning with learning rates of 0.001, 0.001, and 0.00001, respectively.

In addition to this, we have some of the variants of CYN listed below:

CYN Variants: We consider a few variants of CropYieldNet, CYN, to evaluate the significance of its different modules:

BCYN: We call our baseline of CYN as BCYN which does not have a depth-level selection module and contrastive learning. Soil data processed by SDE to learn information across bins is directly input to LSTM. Only one loss MSE is used for learning.
BCYN_SDM: This model utilizes the DSM module exploiting the depth-variant information in soil data with BCYN.
CYN: This is our final model which makes use of all the advancements such as contrastive learning and DSM with BCYN.

3.3.9 Results and Discussion: CropYieldNet

Through the first set of experiments carried over the US counties and Indian districts, given in Table 3.7 and 3.8, respectively, we make the following observation:
• All the methods including baselines, CYN, and its variants show that RMSE is considerably lesser for both the crops when times series length is taken as **padded crop cycle** in comparison to the whole year length. However, this reduction in RMSE is larger for baselines for MODIS data than that of Landsat and Sentinel Data. This is because MODIS data is larger and has better temporal resolution.
• The impact of selecting the appropriate depth for soil attributes (adding SDM to BCYN) overtaking soil attributes at all depths as different attributes is clearly visible in Table 3.7. The maximum % improvement of **BCYN_SDM** over BCYN is achieved for Sentinel-2 (S-2) which is 7.56 and 8.07 for corn and soybean, respectively for the US. Similar patterns are observed for India with an improvement of 4.05% in corn and 13.49% in soybean (Table 3.8).
• It can also be observed from Table 3.7 and 3.8 that our **final model CYN** shows improvement over all the variants and existing models for both crops for all regions. The improvement is maximum for MODIS and minimum for Landsat-8/Sentinel-2. This is because of the temporal resolution of the satellites.
• On **comparing our model CYN** and its variants with the existing models, we observe that CYN outperforms for all crops and regions (US and India) for all satellite data. For example, the RMSE achieved for soybean in the US using Landsat-8 data is 7.370 bu/ac

Table 3.7: RMSE (bu/ac) achieved by baselines, CYN and its variants with yearly and PCC data from different satellites for US

Corn (US)								
	M16		**M7**		**L-8**		**S-2**	
Model	Y	PCC	Y	PCC	Y	PCC	Y	PCC
CNN [43]	30.08	27.169	31.345	28.673	24.619	24.617	27.766	26.997
CNN+GP [43]	29.857	28.204	32.083	30.165	24.365	23.881	31.366	31.049
CNN+LSTM [44]	24.565	23.586	26.926	27.634	23.931	23.632	26.36	25.477
BCYN	24.379	23.069	26.197	25.638	23.296	23.253	24.426	24.356
BCYN_SDM	21.754	21.717	25.854	25.12	21.861	21.84	22.624	22.513
CYN	21.545	21.505	21.595	21.517	21.834	21.819	22.533	22.459
Soy (US)								
	M16		**M7**		**L-8**		**S-2**	
Model	Y	PCC	Y	PCC	Y	PCC	Y	PCC
CNN [43]	14.8	12.436	10.861	10.696	8.679	8.346	10.102	9.679
CNN+GP [43]	10.303	10.115	8.911	8.592	8.561	8.343	8.737	8.383
CNN+LSTM [44]	10.028	9.956	10.472	9.998	8.412	8.37	11.613	10.956
BCYN	9.248	9.212	8.302	8.262	8.446	7.794	8.631	8.372
BCYN_SDM	9.167	9.122	7.973	7.97	7.564	7.476	8.059	7.969
CYN	8.379	8.166	7.898	7.895	7.484	7.37	7.972	7.964

and this is approximately 11% more accurate than all three competing models. CYN obtained RMSE value for soybean in India as 9.840 bu/ac and this is more precise than CNN, CNN+GP, and CNN+LSTM model by 29.79%, 13.75%, and 37.06%, respectively. We can see the same patterns in results for corn in both regions.

Varying number of bins for satellite data: We experiment for observing the behavior of our model on varying the number of bins (B) for surface reflectance data. The experiments are performed for 32, 64, 128, and 256 bins. Table 3.9 shows the results obtained by CYN for corn in the US using Landsat-8 data. It is clear that RMSE achieved for more number of bins makes the model converge faster with marginal improvement in prediction accuracy. The data with more bins will have higher dimensions leading to accuracy improvement. However, the best reduction is obtained when the number of bins is increased to 64 from 32. Similar behavior is observed for different satellite data and for Indian regions. Thus, all the results are taken with histograms of 64 bins.

Table 3.8: RMSE (bu/ac) achieved by baselines, CYN and its variants with PCC data from different satellites for India

	Corn (IN)			**Soy(IN)**		
Model	**M7**	**L-8**	**S-2**	**M7**	**L-8**	**S-2**
CNN [43]	20.03	15.145	17.645	15.219	14.016	14.848
CNN+GP [43]	18.224	15.243	16.736	12.291	11.41	11.837
CNN+LSTM [44]	17.765	15.585	15.64	12.022	15.636	13.37
BCYN	15.445	15.008	15.535	9.817	10.268	11.335
BCYN_SDM	15.386	14.626	14.906	9.816	9.877	9.805
CYN	15.307	14.599	14.852	9.802	9.864	9.795

Table 3.9: RMSE obtained for corn in US using Landsat-8 using histograms with different number of bins

#Bins	RMSE	#Iterations
32	21.9607	32000
64	21.81921	16000
128	21.79979	11000
256	21.78927	7500

Significance of Data Augmentation (DA): Table 3.10 shows RMSE and corresponding % improvement in RMSE with respect to the predictions without DA techniques to see the impact of DA techniques applied on training data from Landsat-8 and Sentinel-2. It is evident from the table that DA has significantly reduced the error in yield prediction. The best results are obtained by CYN when data is augmented using both inversion and temporal irregularity techniques. Data augmentation has worked the best for Sentinel-2 irrespective of crop and location. For example, the best RMSE obtained for corn in India is 11.328 bu/ac on applying both DA techniques on Sentinel-2 data. Data augmentation showed the maximum percentage improvement of 23.2% for Sentinel-2, followed by Landsat-8 with 22.09%, and the least improvement is seen for MODIS with 18.99%. These results also prove our hypothesis that the satellites with high spatial resolution can outperform even with limited data.

Table 3.10: Impact of different Data augmentation techniques on crop yield prediction using CYN. (RMSE in bu/ac) Corresponding %age improvement in RMSE is given in ()

Corn						
	US			INDIA		
	M	L-8	S-2	M	L-8	S-2
No DA	21.517	21.819	22.459	15.370	14.599	14.852
Inverse Time Series	21.505 (0.05)	18.692 (14.33)	19.558 (12.91)	12.483 (18.78)	11.490 (21.29)	11.366 (23.47)
Temporal Irregularity	21.449 (0.32)	18.812 (13.78)	19.676 (12.39)	12.469 (18.87)	11.507 (21.17)	11.475 (22.73)
Both	21.226 (1.35)	**18.587** (14.81)	18.938 **(15.67)**	12.451 (18.99)	11.374 (22.09)	**11.328 (23.72)**
Soybean						
No DA	7.947	7.370	7.964	9.812	9.840	9.795
Inverse Time Series	7.925 (0.27)	6.421 (12.87)	6.190 (22.27)	7.773 (20.78)	7.711 (21.63)	7.679 (21.60)
Temporal Irregularity	7.942 (0.06)	6.425 (12.82)	6.682 (16.09)	7.873 (19.76)	7.732 (21.42)	7.712 (21.26)
Both	7.463 (6.09)	6.210 (15.73)	**5.990 (24.78)**	7.686 (21.66)	7.686 (21.89)	**7.186 (26.63)**

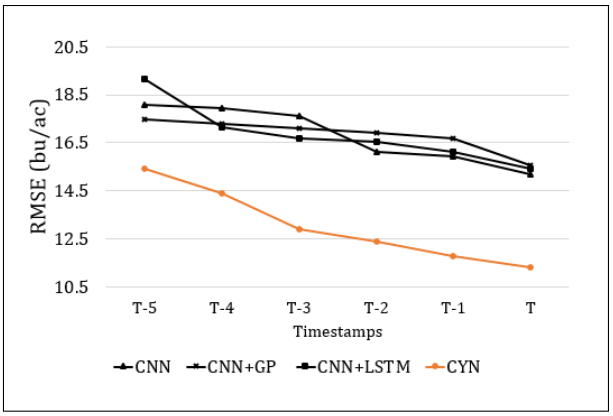

Figure 3.9: Error in In-season Yield Prediction

Table 3.11: Comparison of In-season yield prediction for corn in India by different models using Sentinel-2 data

Model	CNN[43]	CNN+GP[43]	CNN+LSTM[44]	CYN
T-5	18.106	17.488	19.139	15.409
T-4	17.929	17.313	17.135	14.402
T-3	17.597	17.106	16.678	12.892
T-2	16.134	16.93	16.541	12.387
T-1	15.925	16.705	16.124	11.781
T	15.194	15.588	15.408	11.328

In-season prediction: All the results given till now are end-of-season predictions. We have modified CYN to predict yield at multiple in-season timestamps for a crop. Also, we have captured in-season predictions for other existing models by limiting the data input to a smaller number of timestamps. The results for early prediction for corn in India using Sentinel-2 are given in Table 3.11 and the corresponding percentage error is shown in Figure 3.9. As expected, the prediction accuracy improves as we approach towards harvest time. CYN predicts the yield at timestamp T-5 with an error of 0.71% wrt the prediction at timestamp T which is an order lower than 19.17%, 12.18%, and 24.21%, obtained by CNN, CNN+GP, CNN+LSTM, respectively.

Analyzing models for generalization capability: We have analyzed our model for its generalization capabilities to learn from data of different crops or regions. Cross-crop/

cross-region training is required to predict the yield using CYN. The results are given in Table 3.12 for Landsat-8 and Table 3.13 for Sentinel-2.

Table 3.12: Comparison of generalizable capability of CYN with other existing models for Landsat-8
Corresponding %age change in RMSE is given in ()

Soybean (US)			
Model	Inverse TS	Cross-Crop	Cross -region
CNN [43]	8.259	16.092 (48.64)	10.874 (23.99)
CNN+GP [43]	6.342	9.589 (17.97)	9.886 (20.43)
CNN+LSTM [44]	8.103	12.917 (36.13)	10.482 (21.3)
CYN	6.210	6.414 (3.01)	6.353 (2.08)
Soybean (India)			
CNN [43]	13.820	26.379 (47.61)	14.120 (2.12)
CNN+GP [43]	10.527	16.563 (36.44)	10.614 (0.82)
CNN+LSTM [44]	14.012	17.579 (20.29)	14.171 (1.12)
CYN	7.711	8.200 (5.97)	7.769 (0.75)
Corn (US)			
CNN [43]	23.804	28.462 (16.37)	37.243 (36.09)
CNN+GP [43]	22.336	24.755 (9.77)	25.281 (11.65)
CNN+LSTM [44]	23.974	25.387 (5.57)	25.612 (6.4)
CYN	18.692	18.819 (0.68)	18.700 (0.04)
Corn (India)			
CNN [43]	14.933	16.836 (11.31)	25.390 (41.19)
CNN+GP [43]	14.757	16.254 (9.21)	22.133 (33.33)
CNN+LSTM [44]	14.936	17.603 (15.15)	28.076 (46.8)
CYN	11.409	11.490 (0.71)	11.487 (0.68)

Cross-crop training: We combine the data for two crops, corn, and soybean, sharing a similar crop cycle. The major challenge in combining the data for different crops is the different ranges of their target yield. To handle this challenge, we have normalized our target yield using min-max normalization before combining the data and denormalizing it at the prediction end. In order to distinguish between the crops, we add crop as a feature. It is necessary to do this so that anomalies generated can be handled. An anomaly will occur if two crops are grown in the same location (county or district).

Table 3.13: Comparison of generalizable capability of CYN with other existing models for Sentinel-2
Corresponding %age change in RMSE is given in ()

Soybean (US)			
Model	Inverse TS	Cross-Crop	Cross -region
CNN [43]	9.485	23.771 (60.1)	19.842 (52.2)
CNN+GP [43]	8.116	9.884 (17.89)	10.228 (20.65)
CNN+LSTM [44]	9.417	14.750 (36.15)	19.276 (51.15)
CYN	6.190	6.392 (3.17)	6.268 (1.25)
Soybean (India)			
CNN [43]	14.788	33.441 (55.78)	19.623 (24.64)
CNN+GP [43]	11.740	16.978 (30.85)	20.240 (42)
CNN+LSTM [44]	13.169	31.214 (57.81)	15.232 (13.54)
CYN	7.679	8.089 (5.07)	7.690 (0.14)
Corn (US)			
CNN [43]	25.947	35.058 (25.99)	31.633 (17.98)
CNN+GP [43]	27.983	33.009 (15.23)	30.158 (7.21)
CNN+LSTM [44]	25.317	37.332 (32.18)	33.815 (25.13)
CYN	19.356	20.752 (6.73)	19.432 (0.39)
Corn (India)			
CNN [43]	15.677	29.511 (46.88)	31.396 (50.07)
CNN+GP [43]	15.942	25.324 (37.05)	30.905 (48.41)
CNN+LSTM [44]	15.610	19.643 (20.53)	28.260 (44.76)
CYN	11.336	11.738 (3.43)	11.351 (0.13)

Cross-region training: To study the behavior of models when trained on the data from different regions growing the same crop, we use US (county level) and India (district level) data. For cross-region training, the key points that need to be taken care of include the same measuring unit of the target yields, the same duration of the crop cycles, and padding a few timestamps to compensate for differences in crop cycles.

Tables 3.12 and 3.13 clearly show that the RMSE increases for all the models when trained across crops and regions. However, the increase in RMSE is minimal for the proposed model CropYieldNet, CYN. In contrast to this, the existing models drastically fail to precisely predict the crop yield in either of the scenarios. For example, the increase

in RMSE for corn in the US using models trained on both corn and soybean in the US on Landsat-8 data is 16.37%, 9.77%, 5.57%, and 0.68% for CNN, CNN+GP, CNN+LSTM, and CYN, respectively. On the same line, when the models were trained on corn for both the US and India and tested on corn in the US, the corresponding increases in RMSE were 36.09%, 11.65%, 6.40%, and 0.04%, respectively. Similar patterns were observed with Sentinel-2 data.

3.4 Main Contributions

In this chapter, we have focused on analyzing the impact of using conventionally collected data and satellite data and validated the results for the crop yield prediction problem. Also, we have focused on correct and efficient modeling of the CYP problem. The major contributions of the chapter are:

- The proposed model YieldPredictNet (YPN) models CYP as a spatiotemporal problem using simple numeric data and integrates time series with static data.
- We have introduced two modules for handling soil parameters: first an attribute selection unit for soil parameters to minimize the impact of the irrelevant and noisy features on crop yield prediction and second unit to select the appropriate depth for the soil factors that are collected at varying soil depth at different time steps of the crop. These modules are used in both models.
- We have modified the models to perform early *in-season* predictions for a crop yield at multiple stages with comparable accuracy at harvest time prediction.
- We performed extensive comparative analysis on MODIS, Landsat-8, and sentinel-2 data using CYN incorporating data from different modalities, trained on multiple regions, and trained on multiple crops.
- We have also given data augmentation techniques for satellite histogram time series to overcome the problem of data scarcity and more precise prediction of the crop yield.

3.5 Summary

In this chapter, we have worked with three modalities of data viz. meteorological, soil, and surface reflectance bands. For this, we developed two models *YieldPredictNet (YPN)* and *CropYieldNet (CYN)*. YPN works with meteorological and soil data collected in a conventional (not remote sensing) way. CYN uses conventionally collected meteorological data along with satellite-based soil data and surface reflectance bands for CYP. We also tried to appropriately model the crop yield prediction problem as a spatiotemporal problem. our approach consolidates many effective design decisions such as 1) handling spatiality by clustering the locations based on meteorological and soil characteristics; 2) using a padded crop cycle to handle any discrepancy in sowing and harvesting time of the crop at various locations; 3) using data at weekly granularity for regular monitoring of meteorological conditions. We introduced two modules viz. attribute selection module and depth selection module for soil attributes. The attribute selection module selects the appropriate depth-invariant attributes and the depth selection module is used to select the appropriate depth of the soil attribute captured at different levels throughout the crop cycle. Through extensive experimentation, we can make certain recommendations that can prove to be useful for any crop yield prediction system. The recommendation includes the incorporation of as many ”relevant” meteorological attributes as possible. Selecting relevant soil attributes (through FAS) and appropriate depth levels (through SDM) for soil factors plays an important role in improving the prediction accuracy and varies throughout the crop cycle. We also recommend clustering counties to capture spatial patterns. We retrained our model with cluster-specific data to improve the model. The experiments show that the best-suited granularity is weekly and the length of the time series is a padded crop cycle.

The motivation for using recently launched satellites like Landsat-8/9 or Sentinel-2 is their capability to capture high spatial resolution images. However, this leads to the problem of data scarcity which we have addressed using data augmentation techniques. Our extensive experimentation shows the efficacy of our model presented in the chapter.

We also show that when working with satellite data, we can achieve better results even without explicitly handling the spatiality.

The work carried out in this chapter helped us to identify the research gaps more precisely and to decide the future course of the thesis.

Chapter 4

PatchNet: Efficient Representation learning of high-spatial resolution Satellite Image Time Series

[1]

4.1 Introduction

Popular satellite systems which make their data publicly available include Landsat-8/9 [40], Sentinel-2 [41], and MODIS [10]. The last decade has witnessed a significant improvement in sensor technology leading to the availability of higher spatial and temporal resolution satellite images. Many applications like predicting crop yield, snow cover, solar energy, forest fire, etc. require high spatial resolution satellite image time series (HSRSITS) data.

As compared to low spatial resolution satellite image time series (SITS), the amount

[1] *The work presented in this chapter has resulted in the following publication and patent:*

- *Poonam Goyal, Arshveer Kaur, Arvind Ram, and Navneet Goyal, "Efficient Representation Learning of Satellite Image Time Series and Their Fusion for Spatiotemporal Applications", in AAAI Conference on Artificial Intelligence (AAAI 2024).*
- *Provisional Patent Filed: 202411011144 dated 17 February 2024.*

of data associated with HSRSITS increases manifolds, leading to a computing bottleneck. For example, the combined amount of 7 years' data considered for three applications (Crop yield prediction, snow cover prediction, and solar energy prediction) spanning 2000 US counties is approximately 2.1 TB for MODIS (spatial resolution 500m) and 10.0 TB for Landsat 8 (spatial resolution 30m), respectively. The huge amount of data processing required seriously impedes the democratization of the use of satellite image technology for various applications. In this chapter, we address the problem of impractical computational requirements for processing HSRSITS.

We propose PatchNet which learns prominent patterns in HSRSITS by doing a spatial patch-based partial traversal, e.g., $(1/p)th$ spatial processing of SITS using the idea of beam search and attention mechanism for learnable patch selection as shown in Figure 4.1. The learnable patch selection mechanism eliminates the need for full processing of SITS, thereby reducing the amount of processing by a factor of p with some additional overheads and still achieves state-of-the-art results for end tasks. Existing methods deal with the processing challenges by transforming the images into histograms [43, 44, 58, 84]. A few researchers have also tried to transform images into single-value numeric spectral reflectance indices [39, 85–87]. Both these approaches suffer from significant spatial information loss leading to degraded performance.

4.2 Related Work

Satellite data: Satellite systems like AVHRR [88], PlanetScope [89], CartoSat-1 [90], MODIS [10], Landsat-8/9 [40], Sentinel-2 [41], and others are orbiting around the earth and collecting data at varying spatial, temporal and spectral resolutions. AVHRR has a coarse spatial resolution of 1km while PlanetScope, CartoSat-1 have a high resolution of 2-3m but their data is not freely available. Popular satellite systems are MODIS, Landsat-8/9, and Sentinel-2 due to their publicly available data which can be used in different real-world applications like disaster management, urban planning, agriculture, climate studies, etc. MODIS launched in 1999 provides data at a spatial resolution of 250-500m with a revisit time of daily or 8 days depending on the product. Landsat-8/9, launched in

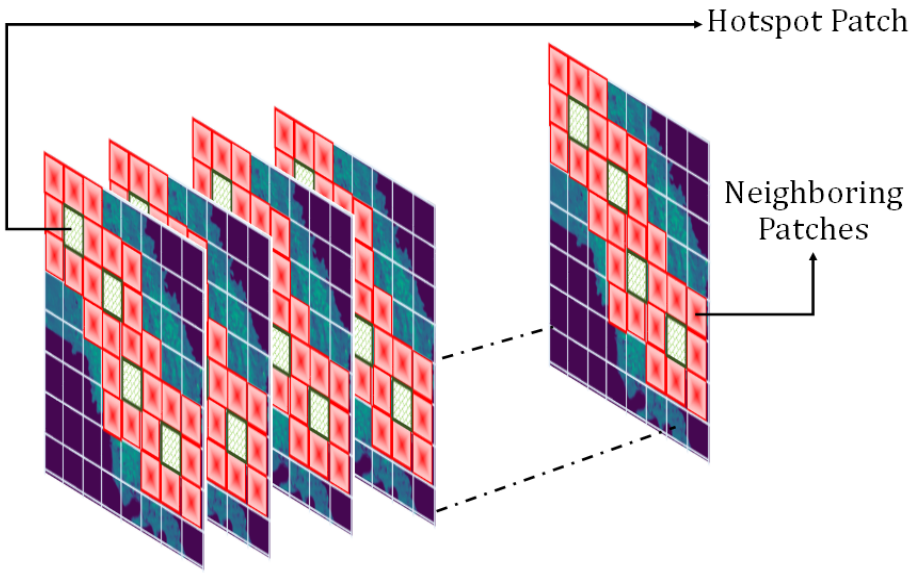

Figure 4.1: Partial Traversal of SITS

2013/2021, has a spatial resolution of 30m with a revisit time of 16 days, and Sentinel-2 launched in 2015 has a spatial resolution of 20m with a revisit time of 10/5 days.

Spatiotemporal Applications: We consider three applications viz. crop yield prediction, snow cover prediction, and solar energy prediction. Accurate crop yield prediction is crucial for ensuring food security around the globe. Researchers have tried to predict crop yield with climate data [59, 64, 91, 92] using traditional machine learning models. These models lack in capturing complex relationships between meteorological attributes and yield. A few researchers applied deep learning models and incorporated genotype [65, 67] and/or soil [44, 93] information. Recent studies attempt to include physics-guided patterns [94], and topological features [95] along with climate data. Research shifted from meteorological data to the use of satellite image data after getting easy access to it. However, it is difficult to process image data due to its high volume. Therefore, vegetation indices are directly computed from MODIS product MOD13Q1 for a location. Researchers [44, 58] and [43] converted MODIS images into histograms and used histogram time series to predict crop yield. Authors [84] presented a deep learning model for MODIS, Landsat, and Sentinel histogram time series to predict crop yield and highlighted the importance of high spatial and high temporal resolution of data required for

the application. However, researchers faced data scarcity for training models using high-resolution satellites launched in the last few years.

The other two applications have gained interest only recently, and very little work is available in the literature. Support vector machine is applied to atmospheric-oceanic dynamics data to predict snow cover. However, satellite technology has a great advantage for collecting data on inaccessible and hazardous regions as compared to proximal sensors and UAV-mounted sensors [96]. Solar energy prediction is done to find a suitable location for the installation of solar plants and reduce the dependence on fossil fuels for economic development. Authors [97] predicted solar energy using Pearson correlation and random forest on meteorological data. Other researchers [98–100] employed climate attributes such as daily minimum and maximum ambient temperature, cloud cover, and day length to predict daily global solar radiation. In recent studies, authors Vico et al. [101], Barrera et al. [102], and Lardizabal et al. [103] employed deep neural networks for daily solar energy prediction using various attributes. The existing studies do not make use of SITS.

4.3 Study Area and Data

We considered the top producers of corn and soybean from the United States for CYP. The crop yield labels are collected from Quick Stats [80] compiled by the United States Department of Agriculture (USDA). For SCP, we have considered the counties which experience average snowfall of more than 250 inches per year. The percentage of the area covered under snow is obtained from the MODIS product MOD10A1 [104]. For SEP, we considered 5 US states.

4.3.1 Data used

We have taken around 2000 counties from different states of US for all applications considered highlighted in Figure 4.2.

Crop Yield Prediction: We have taken top producer counties of corn and soybean from states - Arkansas, Illinois, Indiana, Iowa, Kansas, Kentucky, Michigan, Minnesota, Mississippi, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Tennessee, and

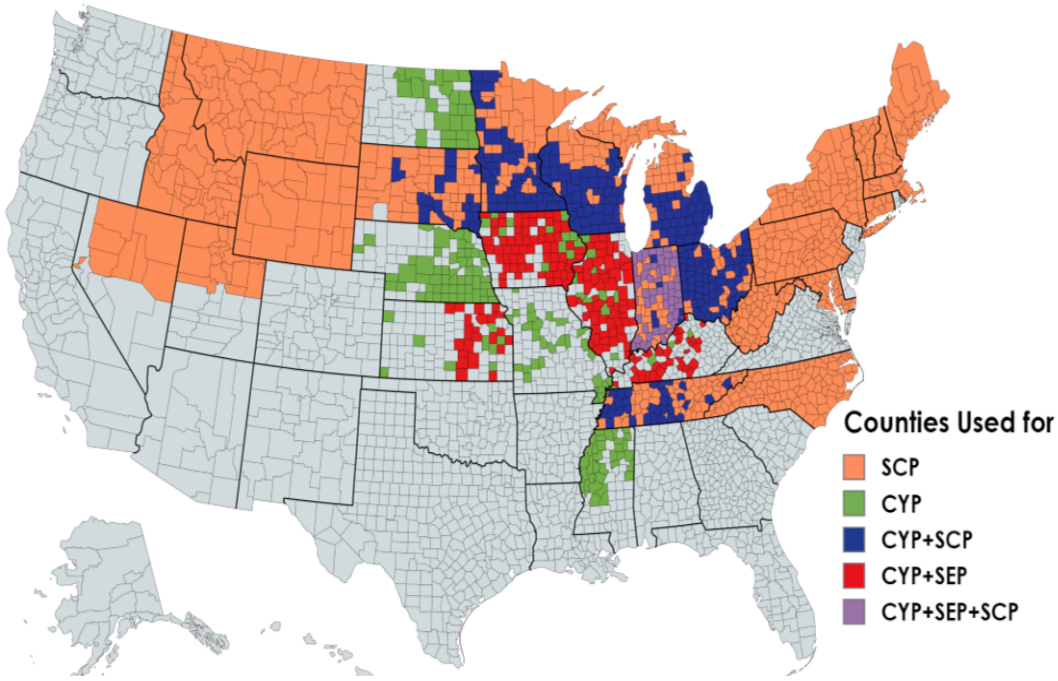

Figure 4.2: Study area for different applications

Wisconsin.

Snow Cover Prediction: We consider 300 counties from states experiencing annual snowfall of more than 250 inches. The counties lie in the states of Washington, Oregon, Utah, California, New Hemisphere, and Colorado.

Solar Energy Prediction: We consider 275 counties from the states- Illinois, Indiana, Iowa, Kansas, and Kentucky.

Meteorological Data: Along with surface reflectance data, we have also used meteorological data as an additional modality that has a direct impact on the considered spatiotemporal applications. Though weather data is available in the form of images through various MODIS products, but the data is available only for four weather attributes - the land surface temperature at night time, precipitation and land surface temperature at day time, vapor pressure, and precipitation. So, we have used meteorological data [105] in a numeric form for 12 attributes collected at a temporal resolution of one day.

Processed Data Volume: The data volume processed for each of the satellites is different

due to their different resolutions. For one MODIS time series for one year is $200 \times 250 \times 5(\#bands) \times 46(\#timestamps)$ pixels which is $\approx$ 150MB whereas for one Landsat-8 time series it is $2000 \times 2000 \times 5(\#bands) \times 23(\#timestamps)$ pixels that makes data volume of $\approx$ 700MB. We have used data for 7 years from 2014-2020. In total this makes, total data processed ≈ 2.1TB for MODIS and ≈ 10TB for Landsat-8. However, we removed the data for counties where the ground truth labels were missing, or if data was not captured by one of the satellites. After performing the entire data cleaning, the final collective data used in three applications is approximately 1.8TB for MODIS and ≈ 9.6TB for Landsat-8.

4.3.2 Data Preparation

Since the satellite data is captured as raw multispectral images. It needs pre-processing before using it for the end task. The data preparation steps are given in Appendix A.

4.4 Problem Formulation

We have considered three spatiotemporal forecasting problems viz. CYP, SCP, and SEP. The goal is to predict $\hat{y}_{c,z}\epsilon$ {crop yield, percentage of area under the snow, and solar energy produced} for a county *c* at prediction time granularity z which is a year, a month and a fortnight for CYP, SCP, and SEP, respectively. Let input data set of TS be $X_z = \{[x_1^1, x_1^2, ..., x_1^t], [x_2^1, x_2^2, ..., x_2^t], ...[x_{z-1}^1, x_{z-1}^2, ..., x_{z-1}^t]\}$, where *t* represents the number of timestamps depending on the application and the satellite. For example, for soybean crop *t*=15 and *t*=30 for Landsat-8 and MODIS, respectively.

4.5 Proposed Model: PatchNet

PatchNet is designed to encode high spatial resolution SITS which is otherwise impractical to process. It works on image times series iteratively for multiple patch time series (patchTS) and uses the idea of a beam search for optimizing the patch selection process. A patch is selected in the spatial dimension and patchTS consists of entire time series for the patch. The architecture of the PatchNet is given in Figure 4.3.

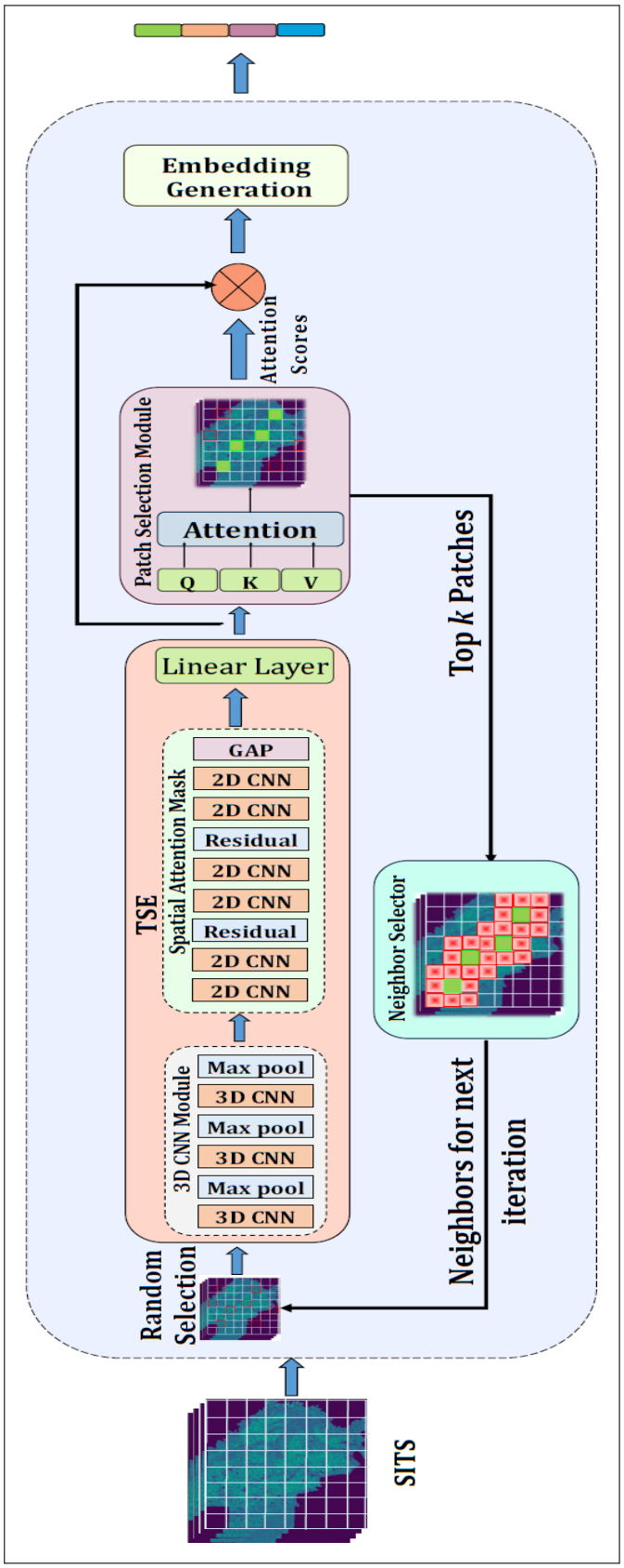

Figure 4.3: PatchNet

We divide the image time series into a spatial virtual grid, resulting in multiple patchTS, one for each cell. We now onwards refer to patchTS as a patch. The patches are processed using TSE and their representations are passed to the Patch selection module (PSM). PSM uses attention score to identify top '*k*' patches that are then forwarded to the Neighbor Selector (NS). NS determines the unprocessed neighboring patches of top '*k*' patches and also creates a list of patches to be processed in the next iteration. The process continues till a fraction (1/*p*) of the SITS is processed. The enhanced patch representations obtained from PSM are passed to the embedding generation module which outputs the embedding of the entire SITS learned by the network in multiple iterations. The pseudo-code of the process is given in Algorithm 1.

Algorithm 1 PatchNet

Input: SITS
Output: Embedding of SITS
initialize: $m = 0\ and$
$|P| = total\ patches\ \ in\ \ SITS$
while $(m) \neq |P|/p$ **do**
  select *n* random patches
  $R = TSE(patchTS)\ \forall\ n$ patches // apply TSE to get linear representation of *n* patches
  $l, \bar{l}, \tilde{R} = PSM(R)$ // list of top *'k'* patches and enhanced patch representations
  $n' = NS(l)$ // *NS* gives neighbors of each patch in *l*
  Select $n - n'$ random patches
  $m = m + n$
  $E_L = EG(\tilde{R})$
return E_L =0

4.5.1 Time Series Encoder (TSE)

TSE gives a linear representation of the input patch. It consists of two submodules 3DCNN network and a Spatial Attention Mask (SAM) followed by a linear layer.

4.5.1.1 3DCNN Module

3DCNN [106] consists of three convolution layers having 10, 15, and 20 filters with zero padding. Each convolution layer is followed by a 3D-max pool layer. 3DCNN leverages both spatial and temporal features simultaneously and learns more informative

representations of the volume.

4.5.1.2 Spatial attention mask (SAM)

We followed [107] and modified it for our problem. It has 6 2D convolution layers, each followed by a batch normalization layer to reduce the internal covariate shift and model overfitting. The skip connections are used after every two convolution layers to improve the information flow within the network and mitigate the problem of vanishing gradient. Global average pooling is done by two pooling operations 'average pooling' and 'max pooling' applied along the channel axis and are concatenated to create an efficient feature descriptor. A convolution layer is applied over the feature descriptor to get the highlighted regions.

4.5.2 Patch Selection Module (PSM)

We utilize the self-attention mechanism (given by equations 4.1 and 4.2) to focus on the most important *k* patches from *n* input patches. PSM learns the enhanced representations of all the patches across iterations using the following process and gives the score of each patch based on its contribution to the end task. The input to PSM is $R = \{r_1, r_2, ...r_n\}$, where *n* is the total number of patches and r_i is the linear representation of each patch after being processed by TSE. The mathematical representation of the mechanism is given below:

Query (*Q*), Key (*K*), and Value (*V*) for self attention are:

$$Q = R \times w_q, \quad K = R \times w_k, \quad V = R \times w_v \tag{4.1}$$

where w_q, w_k, and w_v are weight matrices for *Q*, *K*, and *V*, respectively.

$$A = softmax(QK^T) \tag{4.2}$$

where $A = \{a_1, a_2, ..., a_n\}$ is an attention score matrix for all the *n* patches, and each a_s is of size *b* equal to size of patch embedding.

To get the collective score i.e. contribution of the patch towards the end task is calcu-

lated as:

$$S = \sum_{i=0}^{b} a_s^i \quad s = 1\ to\ n \tag{4.3}$$

$$l,\ \bar{l} = top\_k(S) \quad where\ l + \bar{l} = n \tag{4.4}$$

where top_k is the function that returns a list, *l*, the indices of top *k* patches and a list, $\bar{l}$, remaining patches to be used in the next iteration of the selection process.

PSM also helps in enhancing the patch representations *R* as:

$$\tilde{R} = S \times V \tag{4.5}$$

4.5.3 Neighbor Selector (NS)

NS finds the untraversed neighboring patches of all *k* patches. For a patch p_{ij}, set of neighbors is $\{p_{ef} - p_{ij}\}$, $e = i-1, i, i+1$ and $f = j-1, j, j+1$. Selecting the neighboring patches ensures that focus is maintained near the hotspots and this leverages the geospatial information to boost the prediction. We select a few untraversed random patches for the next iteration to make the number of patches *n*.

4.5.4 Embedding Generation (EG)

EG is a two-level process and consists of two linear layers. The first layer is used to get the representation of each patch, p_{ij} in an iteration. The embeddings of all selected patches across iterations are then concatenated and passed to the second linear layer which gives a representation of the entire SITS.

4.6 Models for Comparison

We compared PatchNet with models which work on histogram time series of satellite data. We considered four existing CYP models CNN [58], CNN+GP [58], CNN+LSTM [44], and CYN [84] working on histogram TS to compare with the proposed PatchNet. A separate histogram is created for each reflectance band by aggregating the pixel intensities of the image into fixed-length bins at every timestamp.

The authors [43] applied CNN and CNN+GP models on histogram time series of reflectance bands for crop yield prediction. They model CYP as a static problem by using CNN models. The models use only surface reflectance data and do not exploit the temporal dependency in the data.

Researchers in [44] used CNN+LSTM over MODIS reflectance band histogram time series and only CNN over soil histograms to model soil as static data, but ignores the depth-sensitive soil information. It models CYP as a temporal problem using soil data and surface reflectance TS. The authors processed raw features using 2DCNN and used LSTM to model the sequence embeddings.

CYN [84] models CYP as a spatiotemporal problem and used soil, and meteorological data along with surface reflectance histograms. It handled the depth-sensitive information of soil as soil data is modeled such that the required depth level is selected for each soil attribute at every timestamp of the crop cycle.

All these models work for different locations, and time duration. However, we used the data for the same locations and time duration in all the models for a fair comparison. To the best of our knowledge, there are no existing models working on histograms for the other two applications.

4.7 Experimental Setup

We performed experiments using Pytorch 1.11.0 and CUDA 11.7 on an A100 GPU server with 80 GB RAM. A model is trained for 50 epochs with a batch size of 8 using Adam optimizer with a learning rate η. We have trained the model with 5 years of data (2014-2018) and, 2 years (2019 and 2020) for testing. To predict the output for the z^{th} year, the training is conducted until the $(z-1)^{th}$ year. For CYP, $\eta = 0.0005$ for a single modality (TSE and PatchNet) and $\eta = 0.000005$ for FuSITSNet. In case of SCP and SEP, $\eta = 0.00001$ for all three models.

4.7.1 Evaluation Metric

We have considered root mean square error (RMSE) for the performance evaluation of the proposed model for the three applications. RMSE measures the error's root mean squared magnitude and penalizes the larger errors compared to the small magnitude errors. The higher RMSE depicts that the model encounters many large-magnitude errors. The formula for RMSE is given as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_{c,z} - \hat{y}_{c,z}^{i})^2}{N}} \tag{4.6}$$

where *N* is the total number of $\{c, z\}$ pairs, *c* is the county, *z* is year, month, or fortnight for CYP, SCP, or SEP, respectively.

4.8 Results and Discussion

Significance of using SITS over histograms time series: The first set of experiments are conducted to compare the proposed model PatchNet with existing CYP models using histogram TS. Table 4.1 presents RMSE (in bu/ac) achieved for corn and soybean yield prediction using various models. It can be observed from the table that for corn yield prediction RMSE reduced by $\approx 12\%$ and 10% with that of CNN and CNN+GP, respectively. These two models use only surface reflectance histograms. The reduction in RMSE is 9% for the CNN+LSTM model which also incorporates meteorological data. CYN uses both meteorological and soil data along with surface reflectance histograms. PatchNet outperforms CYN even without using any other data. However, the error is reduced by an additional 6% when meteorological data is incorporated into PatchNet.

Significance of Number of bands: To create the baseline, we conducted a few preliminary experiments using the histogram data to analyze the impact of bands being taken. The RMSE obtained using all the bands of Landsat-8 time series for corn yield prediction is 24.064bu/ac, and using 5 bands common in almost all satellite systems, RMSE

Table 4.1: Comparison: PatchNet vs histogram models

Model	Corn	Soybean
CNN[43]	24.617	8.346
CNN+GP[43]	23.881	8.343
CNN+LSTM*[44]	23.632	8.370
CYN**[84]	21.819	7.370
PatchNet	**21.469**	**7.290**
PatchNet+M	**20.631**	**6.963**

*uses additional soil data ** uses both meteorological & soil

obtained is 24.3421bu/ac. A similar pattern is observed in other cases as well. There is not much difference in the error, when we carried out the experiments using only 5 bands i.e. Red, Blue, Green, NIR, and SWIR. These bands contribute the maximum in detecting vegetation and various other land covers. The spectral reflectance indices or commonly known as vegetation indices used for monitoring agriculture, water stress etc. are all derived from only these 5 bands. So, we can say that these five bands are sufficient for the selected applications. So, we have taken only common bands i.e. Red, Blue, Green, NIR, and SWIR for all the experiments.

Significance of patch selection mechanism: We conducted experiments without using PSM and NS in PatchNet and instead replaced them with random patch selection. RMSE achieved by random selection is 24.29 bu/ac and 9.98 bu/ac for corn and soybean yield prediction in comparison to 21.47 bu/ac and 7.29 bu/ac, respectively using PatchNet. It is evident that there is a significant improvement in the model performance and PSM and NS collectively work effectively to exploit the required hot spot features in the SITS and eliminate the need to fully process it. Also, it suppresses the noise in the two modalities.

Significance of using Bits Precision: We tested the pipeline used in PatchNet based on a random selection of n patches. RMSE (in bu/ac) obtained using float SITS is 24.764 and 10.416 for corn and soybean, respectively. In the case of using unsigned integer SITS, the respective RMSE achieved is 24.912 and 10.514 bu/ac. The change in RMSE is less than 1% in both cases. So, we converted the entire dataset into unsigned integer images

as it reduced the storage space significantly. All the results presented in this chapter are performed on unsigned integer SITS.

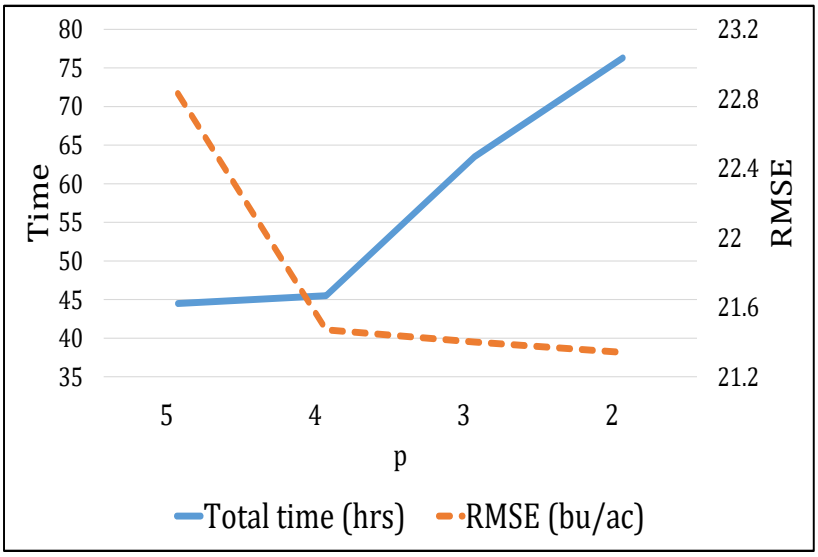

Figure 4.4: Deciding p for (1/p)th traversal of SITS

Deciding (1/p)th traversal of SITS: The next set of experiments is performed for PatchNet to know the optimum number of patches for the traversal of Landsat-8 image time series. Figure 4.4 shows the computation time required and RMSE curve for corn yield prediction by varying p as 5,4,3, and 2 keeping the patch size of $H \times H, H = 64$. It can be observed from Figure 3 that, there is an improvement of $\approx 6\%$ in the performance of the model when the traversed region p changes from 5 to 4, and RMSE did not change much after that. However but the computation time required from p = 5 to 4 is almost constant but it increases linearly after that.

We also experimented by changing the **patch size** to $H = 128$ for $p = 4$ and found that RMSE reduced to 21.42 bu/ac which is just 0.2% as compared to that with patch size 64, but the computation time required to process the patch size of 128 increased 1.5 times. To maintain the trade-off between the computation time and RMSE, we chose to carry out the results by taking $p = 4$ and patch size $H = 64$.

4.9 Main Contributions

The main contributions of the chapter are as follows:

- We introduce a novel model to efficiently process time series of high spatial resolution satellite images. We propose PatchNet which only needs to partially process the image time series using the concept of patches. The patch selection mechanism recommends the most informative patches and achieves state-of-the-art results for the end tasks considered.
- The proposed approach helps in the democratization of satellite technology.

4.10 Summary

We proposed a model to efficiently process the high spatial resolution satellite image time series for earth observation applications. PatchNet is able to get the representation of SITS by partial traversal. The learnable mechanism of patch selection has shown superiority over a random selection of patches to get the SITS representation.

The experimentation shows that PatchNet outperforms the existing models working with histogram time series of satellite data. By preserving spatial information and temporal dynamics, using image time series provides a richer representation of changes in the Earth's surface over time. This enables finer-grained analysis, facilitating precise identification of trends, patterns, anomalies, and more accurate monitoring of different earth observation applications.

Chapter 5

Fusion of two Satellite Image Time Series: Best of both worlds representation learning

[1]

5.1 Introduction

Satellite systems are characterized by their varying spatial, temporal, and spectral resolutions. MODIS has a moderate spatial resolution, which varies between 250 meters to 1 kilometer, and a revisit time of one day or 8 days. Despite having a satisfactory temporal resolution, the coarse spatial resolution of MODIS renders it either unsuitable for applications like urban planning and precision agriculture or imprecise for applications like crop yield prediction, snow cover, etc. [108]. The later applications require data

[1] *The work presented in this chapter has resulted in the following publications:*

- *Arshveer Kaur, Poonam Goyal, and Navneet Goyal, "LSFuseNet: Dual-Fusion of Landsat-8 and Sentinel-2 Multispectral Time Series for Permutation Invariant Applications", in IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA 2023).*
- *Poonam Goyal, Arshveer Kaur, Arvind Ram, and Navneet Goyal, "Efficient Representation Learning of Satellite Image Time Series and Their Fusion for Spatiotemporal Applications", in AAAI Conference on Artificial Intelligence (AAAI 2024).*

from higher-spatial resolution satellites. Landsat-8 and Sentinel-2 have a spatial resolution of 30m and 20m, with a revisit time of 16 and 10 days, respectively. Although these satellites have a finer spatial resolution, their usability is often constrained by various atmospheric effects, like clouds and shadows, which makes their temporal resolution irregular and coarser. This increases the imprecision for applications such as the prediction of crop yield, snow cover, etc. All these satellites either have a high spatial resolution or finer temporal resolution but cannot achieve both simultaneously. Due to budgetary and technological constraints, it is still not possible to capture satellite images with required high spatial and temporal resolutions using a single publicly available satellite system [109]. This necessitates the development of efficient fusion algorithms that combine high spatial resolution images of one satellite system (with low temporal resolution) with high temporal resolution images of another satellite system (with low spatial resolution), and vice-versa.

NASA and the US Geological Survey (USGS) collaborated to create the Harmonized Landsat Sentinel-2 (HLS) data product. HLS merges Landsat-8 and Sentinel-2 data to offer reliable, and seamless data [110]. By utilizing reflectance bands from the two satellites, temporal resolution can be significantly enhanced to 5 days. The spectral resolution of the data is the same as that of Sentinel-2. However, the difference in spatial resolution between the two datasets is not optimally exploited because the HLS data is generated using the resampling approach proposed in [111]. As a result, HLS data provides high temporal resolution with a lower spatial resolution of 30 meters [112].

HLS data has opened up the possibility of fusing data from different satellites to achieve better resolutions. Data can be fused in many ways– pixel-level, feature-level, and decision-level [113]. The pixel-level fusion requires mapping the corresponding pixels in two images e.g. STARFM [27] which fuses the images obtained from MODIS and Landsat-8. The limitation of this fusion technique is that it requires at least one pair of images captured on the same day which is not possible in the case of Landsat-Sentinel fusion. A few attempts are made to fuse the data from different sensors of the same satellite [28–31] to produce high-spatial and spectral resolution imagery. A study uses a

generative adversarial network (GAN) based super-resolution method to reconstruct the synthetic Landsat images to match the spatial resolution of Sentinel-2 images [114]. Also, performing the pixel-level fusion of these high spatial resolution satellites adds up another challenge of storing and processing huge sizes of multi-spectral images. These models are constrained by the requirement of availability of Landsat and MODIS images captured on the same day and may also propagate existing noise. Moreover, the generation process is slow. If we use the generative models to enhance the temporal resolution of SITS, it will increase the amount of data twofold and would thus be computationally prohibitive. The decision-level fusion entails integrating the extracted features from both satellites to arrive at a single decision using voting or weighting methods [113]. The feature-level fusion combines the features from both satellites to perform the end task. However, in the name of feature-level fusion, the existing models [115, 116] have taken spectral reflectance indices like NDVI, GNDVI, NDWI, etc. from two satellites to perform the end task. Feature-level fusion allows us to handle the bottleneck of the two satellites having completely different visiting days, and varying spatial and spectral resolutions.

In this chapter, we propose two fusion models- LSFuseNet and FuSITSNet. LSFuseNet performs a feature-level fusion of histogram time series of Landsat-8 and Sentinel-2. FuSITSNet is a twofold feature-based fusion model which can be used to fuse any two satellite image time series. We have applied it for Landsat-8 and MODIS SITS.

5.2 Related Work

To handle the resolution trade-off of satellite systems, spatiotemporal fusion is one of the possible solutions. Fusion of the satellite data can be done at three different levels– pixel level, decision level, and feature level.

Pixel-level fusion: STARFM [27] creates synthetic Landsat-like image at timestamp $t + 1$ by fusing MODIS and Landsat images at time t. It is a linear model which selects the neighboring pixels with similar spectral properties at the fine resolution and calculates a weighted sum to calculate the final predicted reflectance values. STARFM is a strictly pixel-based method and struggles to achieve satisfactory results in heterogeneous land-

scapes and needs at least one pair of images captured on the same day. Other variants of the method have also been developed but they also suffer from similar challenges. Another study uses a linear regression model on pixels to generate images [117]. NASA, in its HLS data project, fused Landsat-8 and Sentinel-2 to create temporally denser images with a spatial resolution of 30 meters. A model is introduced for the fusion of common bands from Landsat-8 OLI and Sentinel-2 data using the upscaling and downscaling principle along with a regression model [111]. Shao et.al. fused Landsat-8 and Sentinel-2 data using the super-resolution convolutional neural network at each band individually. Wu et.al. [109] proposed an approach to simulate the coarse-resolution image on the reference date with the help of a fine-resolution image. The approach was able to give satisfactory results for MODIS-Landsat fusion, but the performance of the approach was not up to the mark for Landsat-Sentinel fusion. Generative Adversarial Network-based super-resolution method is used to predict the synthetic Landsat images having a spatial resolution as that of Sentinel-2 images [114]. All these studies focused on the pixel-level fusion of the high-resolution images of Landsat-8 and Sentinel-2 which limits their applicability on a large scale due to high memory and computational requirements. The pixel-based methods blindly use noisy pixels in the fusion process, thus propagating the noise in the neighboring pixels of the predicted image [118].

Decision-level fusion: The authors in [113] use the decision-level fusion method for fusing the data from Sentinel-1 and Sentinel-2. The study focused on explicitly selecting the features and observing their impact on crop type classification.

Generative Models: Given the limitations of pixel-based methods, learning-based approaches are gaining interest due to their flexibility and adaptability in capturing complex relationships from the data without relying on predefined assumptions, like in STARFM. Authors in [119] and [120] used downscaling and upscaling approaches to generate an image having a spatial resolution of Landsat-8 with the help of a MODIS image. A few attempts have been made to use advanced Generative adversarial networks (GAN) for image generation. The model in [121] generates a Landsat image at time t using MODIS and Landsat images at t and $t-1$, respectively. Similarly, [118] also used GAN to handle

noise while generating a Landsat-like image using a MODIS image at timestamp t and two Landsat images at $t-1$ and $t+1$. Because of the complex image generation process, generative models have been applied to small-scale datasets having only a few locations are considered for the study [121]. The applications under consideration require time series at a high temporal resolution and interpolating images between two consecutive images is inefficient and computationally expensive. Moreover, this approach increases the data volume twofold when we generate images at mid-timestamps. Also, the generated images can increase the already existing noise in the original images.

Feature-level fusion: To overcome the computational problem, researchers tried to fuse the data from two satellites by retrieving various spectral reflectance indices making it computationally easy to work with, and named this feature-level fusion. Schreier Jonas et. al [122] fused Landsat and Sentinel data with the help of MODIS data to perform the crop-specific phenomapping using NDVI values. Another study [115] focuses on integrating the data from Landsat-8 and Sentinel-2 for yield prediction of sugarcane using green normalized difference vegetation index (GNDVI) obtained from Landsat-8 and Sentinel-2. These methods were applicable to only specific applications as all the vegetation indices are not applicable to all the applications.

Few studies have been carried out using HLS data. Pastick Neal et.al. [110] proposed a regression approach for monitoring the land surface phenology. HLS data derived Normalized Difference Vegetation Index (NDVI) values are compared with the NDVI values captured using MODIS to evaluate the accuracy of the model. Similarly, another study focuses on predicting wheat yield using the vegetation indices obtained from HLS data [108]. Griffiths Patrick et.al. used a few red edge bands from HLS data for crop and land cover mapping [74]. Although, HLS has the advantage of the harmonization of data and the consistency in spectral bands as if the data were acquired by a single sensor [108]. But, HLS is more susceptible to cloud cover as compared to Landsat-8 and Sentinel-2 individually, and HLS data is not able to use the high resolution of Sentinel-2 as its spatial resolution is limited to 30m.

In this chapter, we introduced two models *LSFuseNet* and *FuSITSNet* as a solution

to overcome the limitations mentioned above. The models fuse the time series of two satellites and can learn the features from the time series of varying spatial, temporal, and spectral resolutions. The models incorporate advancements in multi-modal learning and integration of heterogeneous data used in NLP/vision [123–125] to address the limitations of using data from a single satellite and handle the challenges encountered while using multi-modality data from two different satellites. We hypothesize that if the data from the two satellites are modeled and fused appropriately it can achieve comparable results without any additional data. The same is evident from the results of our proposed models. Still, we have also modified the models to incorporate meteorological and soil data to further enhance the performance.

5.3 Study Area and Data

We have used different data for the two models. LSFuseNet works only for histogram time series and FuSITSNet works for satellite image time series data. The details of study area considered for different applications is given in Table 5.1.

Satellite Data: Surface reflectance bands were acquired from two satellites, Landsat-8 and Sentinel-2 for LSFuseNet. FuSITSNet uses images acquired from two satellites, Landsat-8 and MODIS.

Table 5.1: Study area for different applications

Application	States	Total No. of counties
CYP	Arkansas, Illinois, Indiana, Iowa, Kansas, Kentucky, Michigan, Minnesota, Mississippi, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Tennessee, and Wisconsin	900
SCP	Washington, Oregon, Utah, California, New Hemisphere, and Colorado	300
SEP	Illinois, Indiana, Iowa, Kansas, and Kentucky	275

Yield data: The crop yield data for US counties is obtained from Quick Stats [80], a database compiled by the United States Department of Agriculture (USDA). We have taken top producer counties of corn and soybean from the states given in Table 5.1. The

data span is from 2016 to 2020 for LSFuseNet and 2014 to 2020 for FuSITSNet. The yield values are in bushels per acre (bu/ac).

Snow Cover Prediction: The snow cover data at the county level is collected using MODIS product MOD10A1 [104]. It provides a snow cover extent at a spatial resolution of 500 meters. The algorithm utilized in this approach for identifying snow cover extent is based on the normalized difference snow index (NDSI), which leverages the disparity in reflectance values between snow and non-snow surfaces in the visible and near-infrared spectral regions. NDSI indicates the percentage of area covered under snow. We consider 300 counties from states experiencing annual snowfall of more than 250 inches.

Solar Energy Prediction: The information about solar energy produced in a county has been acquired from [105]. The value represents the total solar energy produced in MJ/m2 for a county in a day. The data used in this study spans the period from 2014 to 2020.

Meteorological Data: Along with surface reflectance data, we have also used meteorological data as an additional modality that has a direct impact on the considered spatiotemporal applications. Though weather data is available in the form of images through various MODIS products, but the data is available only for four weather attributes - the land surface temperature at night time, precipitation and land surface temperature at day time, vapor pressure, and precipitation. So, we have used meteorological data in a numeric form for 12 attributes collected at a temporal resolution of one day.

Soil Data: The soil properties typically remain constant over time at a particular location, which makes soil data independent of temporal resolution. The soil data provides information on various soil properties, such as carbon content, pH in water, clay content, bulk density, water, and sand content. The data is collected at six different depth levels, starting from ground level up to 200 cm. Soil data is applicable only for CYP application and it is not relevant to SCP or SEP.

5.3.1 Data Preparation

Since the satellite data is captured as raw multispectral images. It needs data pre-processed before it can be used for the end task. The data preparation steps followed for LSFuseNet are given:

- *Missing values in satellite data:* The missing values are handled using interpolation methods (see Chapter 2 section 2.4.1).
- *Preparing meteorological data:* The meteorological data is captured at a daily granularity and missing values are estimated using the forward fill method as given in Chapter 2 section 2.6.1.
- *Generating Histograms:* The images obtained from Landsat-8 and Sentinel-2 are converted into separate histogram time series using the process given in Chapter 2 section 2.5.2.1.

FuSITSNet works with image time series and thus needs different preparation steps. The steps followed for FuSITSNet are:

- *Handling cloudy pixels:* The cloudy pixels in the images are handled as given in Chapter 2 section 2.4.1.
- *Bits Precision:* By default the Landsat-8 images have float values at every pixel for all the reflectance bands. The float values are converted into unsigned integers to reduce the storage space required. Bits Precision is applied only to Landsat-8 images and the details are mentioned in Chapter 2 section 2.5.1.1.
- *Number of bands:* As mentioned in Chapter 4 section 4.8 there was not much difference in the error in the preliminary results when all the bands were used and when only 5 bands were used. Thus, we have used only the 5 bands common in both Landsat-8 and MODIS satellite systems i.e. Red, Blue, Green, NIR, and SWIR. Another reason for the same is to have a fair comparison with the generative fusion models, as they are capable of working only with the common bands.

- *Time series length in each application:* The length of the time series for each application is decided as given in Chapter 4 section 4.3.2.2.

5.4 Proposed Model: LSFuseNet

In this section, we introduce our proposed model, LSFuseNet. The broad model pipeline is given in Figure 5.1. First, we have given a brief overview of the technique to effectively fuse the two modalities (Landsat-8 and Sentinel-2) using LSFuseNet and subsequently elaborated on each module of the model.

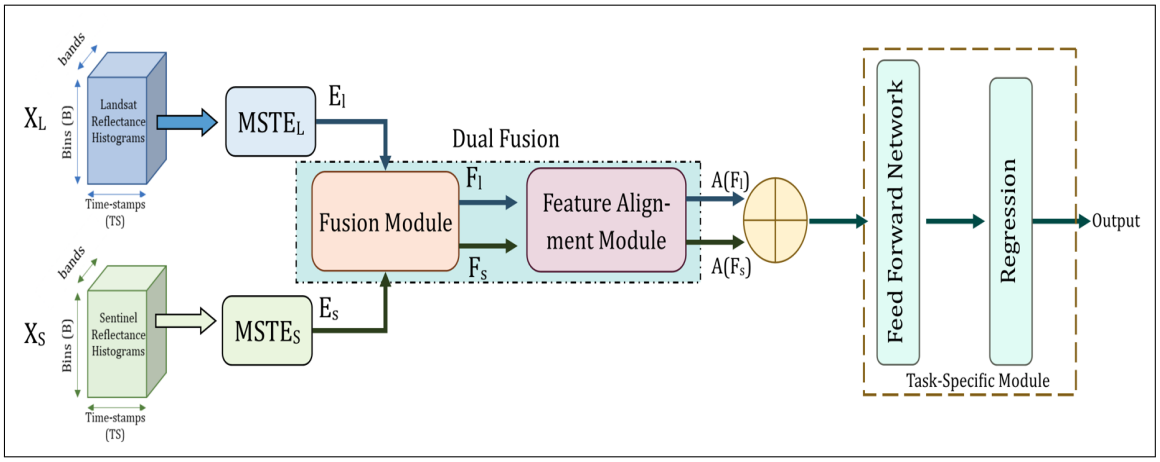

Figure 5.1: Model Architecture: LSFuseNet

5.4.1 Model Overview

The fusion of Landsat-8 and Sentinel-2 using the images has high computational and memory requirements. To utilize the maximum possible information from the satellite data with lesser computational cost, we chose to convert the images into histograms and use these histograms in the fusion technique. LSFuseNet consists of four modules viz. Multispectral Spatiotemporal Encoders (MSTE), Fusion Module (FM), Feature Alignment Module (FAM), and Task-Specific Module (TSM). Two parallel Multispectral Spatiotemporal Encoders (MSTE) learn the individual spatial, temporal, and spectral patterns in surface reflectance data from Landsat-8 and Sentinel-2, respectively. The proposed model uses a novel dual-fusion using Fusion Module (FM) and Feature Alignment Module (FAM). FM and FAM mutually reinforce each other. FM emphasizes and learns the

cross-modal features from both the time series and reduces heterogeneity in the extracted features. FAM handles any noise incurred in cross-modal learning and aligns the features of one modality guided by the other to effectively learn the fine-grained features. Finally, a Task-Specific Module (TSM) is applied to the combined features using a feed-forward network and regression layer to accurately predict the target output.

MSTE is first pre-trained on a larger satellite dataset for two classification tasks to enhance the model performance. MSTE is fine-tuned during the end-to-end learning of the whole model architecture using stochastic gradient descent and contrastive learning.

Input: The input to the model is a histogram time series for a specific location from both satellites. Say, input is represented as X which comprises of two time series X_L and X_S for Landsat-8 and Sentinel-2, respectively. The length of the time series varies depending on the application and the temporal granularity of the satellite. For crop yield prediction, it is the padded crop cycle, and for snow cover prediction, we use three months to forecast the snow for the next month. The histograms for all *TS* timestamps are arranged in columns to form a 2D matrix with dimensions $B \times TS$ for each band, where the band information is considered as depth. The resulting data tensor has a size of $B \times TS \times D$.

5.4.2 Multispectral Spatiotemporal Encoder (MSTE)

The module learns the heterogeneous spatial, temporal, and spectral patterns present in the surface reflectance data of the two modalities. The input data is passed through a 1D CNN encoder that captures the spatial patterns. The encoder consists of 3 convolutional layers with 12, 15, and 20 filters respectively, followed by a maxpool layer after each convolution layer. A bidirectional LSTM layer is applied to learn temporal patterns, with a dropout layer to prevent overfitting. We have used multi-head self-attention to highlight the hotspots in the time series [126]. The attention layer determines weights for the values by mapping query(q) and key-value (k-v) pairs. Multi-head attention divides input and computes it in parallel on different heads, resulting in multiple representations for each (query, key, value) combination. The process reduces the computational cost from that of a single head due to the reduced dimensions at each head. A linear layer is applied

to flatten the tensor of each timestamp, resulting in a 2D output. We have used two encoders in parallel– $MSTE_L$ for Landsat-8 and $MSTE_S$ for Sentinel-2. The output of the respective encoders is represented as E_l and E_s, combined we call it E. The pictorial representation of the module is given in Figure 5.2.

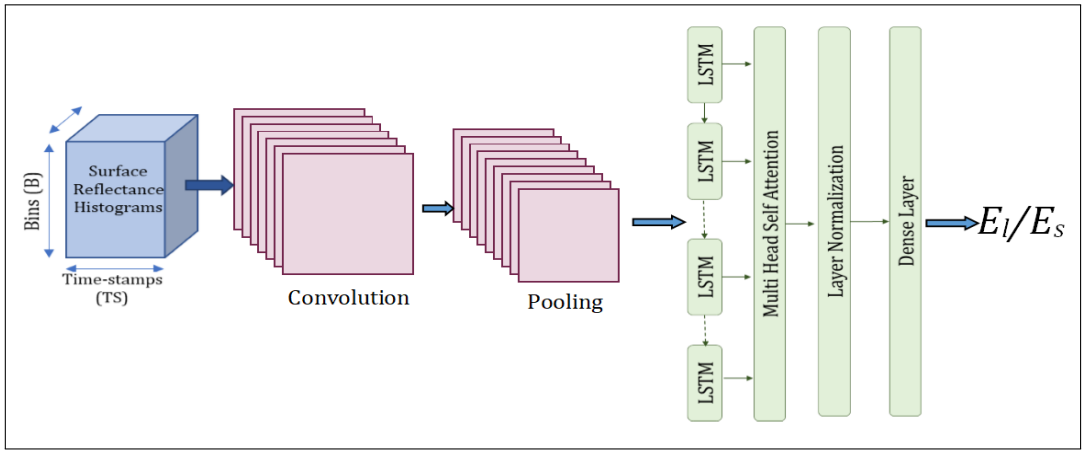

Figure 5.2: Multispectral Spatiotemporal Encoder

5.4.3 Fusion Module (FM)

Fusion Module is used to learn the inter-modality relationships from the two embeddings E_l and E_s obtained from the two pre-trained encoders $MSTE_L$ and $MSTE_S$. The standard attention mechanisms [127] use one modality as the query and compute the cross-attention score. We have used bi-directional cross-modal attention to learn the enhanced features from both modalities. The attention mechanism uses both embedding representations as queries and uses two cross-modal attention layers to learn the relationship between the two modalities - first from Landsat-8 to Sentinel-2 and second from Sentinel-2 to Landsat-8. This helps the module capture the complementary aspects of the two modalities and leverage information from one modality to correct or compensate for the poor-quality data in the other modality. The problem of poor quality data can be more often in satellite data which is a matter of concern. The module outputs two vectors, say F_l and F_s for Landsat-8 and Sentinel-2, respectively. The pictorial representation of the module is given in Figure 5.3.

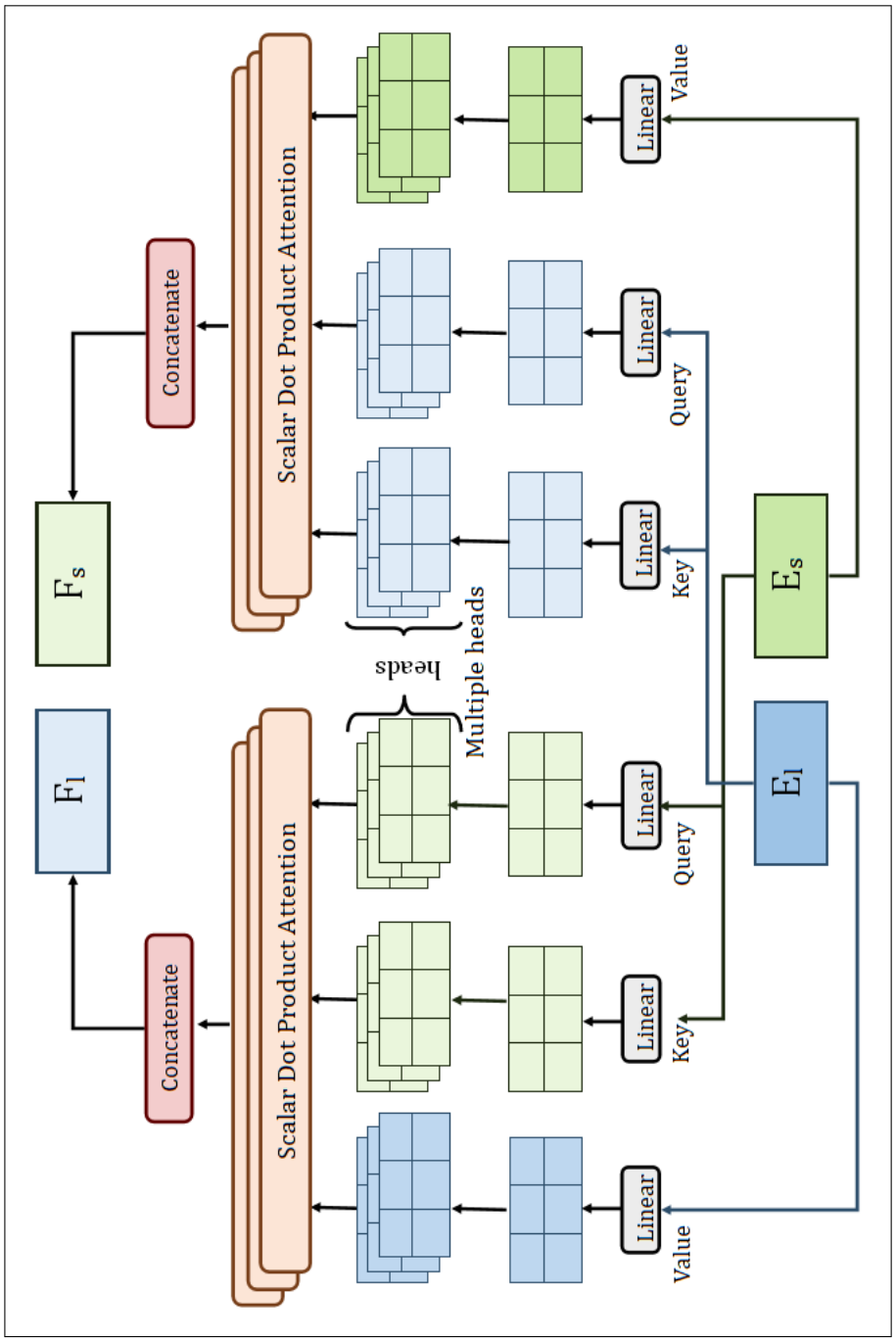

Figure 5.3: Fusion Module:LSFuseNet

5.4.4 Feature Alignment Module (FAM)

The cross-modal attention in FM may induce some amount of noise in the features because of the mismatching of the context (as a result of poor quality data from one/both satellites) of the two multi-spectral time series. We introduced a feature alignment module (used in text video hybrid fusion [124]) to ensure that the features extracted from the two modalities are aligned towards the same context. The feature alignment module captures fine-grained underlying relationships between them. The module performs the feature-wise cross-modal interaction between the two modalities. The sequence length of the Landsat-8 and Sentinel-2 vectors is denoted as *m* and *n*, respectively. We compute the similarity of each feature of Landsat-8 with all features of Sentinel-2 and vice-versa, using the following formula:

$$Sim_L = F_l(F_s)^T \tag{5.1}$$

$$Sim_S = F_s(F_l)^T \tag{5.2}$$

where Sim_L and Sim_S are the similarity scores of Landsat-8 and Sentinel-2, respectively. F_l and F_s are the respective embeddings for Landsat-8 and Sentinel-2 obtained from FM.

We then applied the softmax function over the similarity matrices and used the average feature-wise aggregator over the Sentinel-2 features as shown below:

$$A_i(F_l) = softmax(Sim_S)F_l, \;\; (1 \leqslant i < n) \tag{5.3}$$

$$A(F_l) = [A_1(F_l); A_2(F_l); \cdots ; A_n(F_l)] \tag{5.4}$$

where $F_i(F_l)$ is the similarity-aware aggregated Sentinel-2 representation of i^{th} feature of Landsat-8.

Similarly, we obtained the similarity-aware aggregated Landsat-8 representation of i^{th} feature of Sentinel-2:

$$A_i(F_s) = softmax(Sim_L)F_s, \;\; (1 \leqslant i < m) \tag{5.5}$$

$$A(F_s) = [A_1(F_s); A_2(F_s); \cdots ; A_m(F_s)] \tag{5.6}$$

5.4.5 Task-Specific Module (TSM)

This module implements the end task e.g. regression and consists of three layers of linear transformations with the Gaussian Error Linear Unit (GELU) activation function, and each linear layer is followed by layer normalization. Finally, regression is applied to get the output of the end task.

5.4.6 Soil Data Encoder (SDE)

The Soil Data Encoder (similar to one in Chapter 3 Section 3.6.2) is an additional module that is used to extract features from soil histograms. This module encodes soil information using a CNN as its backbone. The soil information is presented as a 3D tensor in the form of rearranged soil histograms with dimensions $B_s \times 1 \times D_s$, where B_s represents the number of bins and D_s represents the number of depth levels for each soil attribute. The collective input for all attributes is presented as $B_s \times A \times D_s$, where *A* denotes the number of soil attributes. The output of the SDE is a 3D tensor, represented by $\widetilde{SD}$, with dimensions of $B_s^{'} \times A \times D_s^{'}$.

5.5 Pre-training

Taking the motivation from applications in NLP, we pre-trained our encoder with individual data for Landsat-8 and Sentinel-2. While pre-training a model, we took care of- i) the size of the dataset should be sufficiently large to offer the required variation in data patterns and ii) the selection of pre-training objectives. The dataset used for pertaining is $\approx$ 0.1M instances which makes a dataset size of around 1GB for Landsat-8 and 2GB for Sentinel-2. We have pre-trained the encoder (MSTE) for two binary self-supervised classification tasks.

Is reversed (Y/N): During training, the model is presented with time-series data in reversed order along with the original data, and it must predict whether the series has been reversed or not.

Is irregular (Y/N): Irregular time-series data is characterized by inconsistent time intervals. The model is given as input the original regular time series and irregular time series by randomly removing 15% of the original timestamps of a time series.

The model is pre-trained with Binary Cross entropy which loss heavily penalizes for incorrect predictions.

5.6 Proposed Model: FuSITSNet

FuSITSNet (Figure 5.4) consists of two encoders TSE and PatchNet and a fusion module. We use FuSITSNet for fusing two SITS from Landsat-8 and MODIS. We processed Landsat SITS using PatchNet (see Chapter 4). MODIS has a coarser spatial resolution and can be processed as a whole, thus we used TSE for processing its time series. However, we can replace TSE with PatchNet to generate embeddings if the second SITS also has a high spatial resolution.

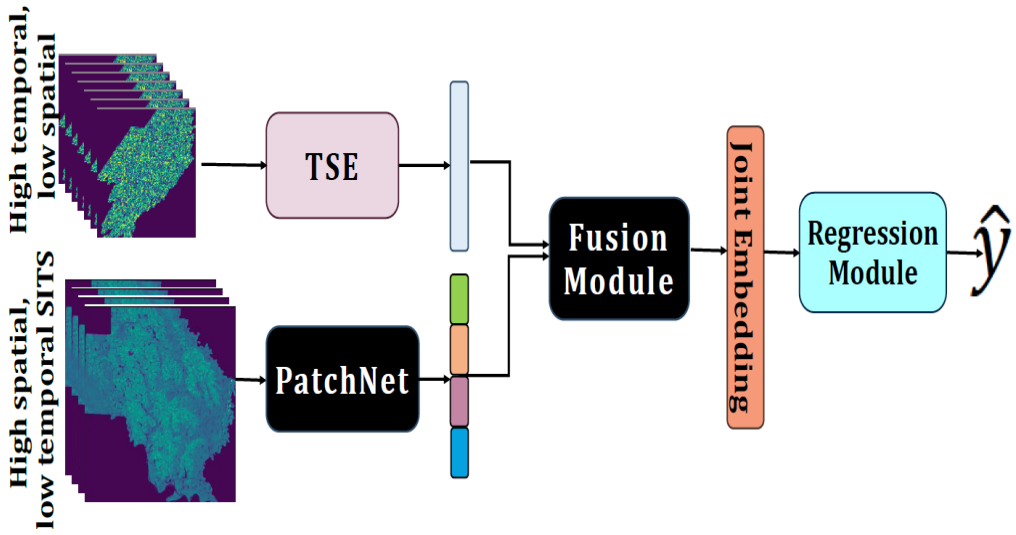

Figure 5.4: FuSITSNet

5.6.1 Time Series Encoder (TSE)

Time Series Encoder (TSE) gives a linear representation of the input patch. It consists of two submodules - 3DCNN network and a Spatial Attention Mask (SAM) followed by a linear layer. Details are given in Chapter 4 section 4.5.1.

5.6.2 PatchNet

PatchNet is designed to encode high spatial resolution SITS which is otherwise impractical to process. It works on image times series iteratively for multiple patch time series (patchTS) and uses the idea of a beam search to optimize the patch selection process. The complete working of PatchNet is given in Chapter 4 section 4.5.

5.6.2.1 Fusion Module

Fusion Module given in Figure 5.5 is a twofold module that takes the embeddings E_M and E_L from the two encoders for MODIS and Landsat-8, respectively. It learns the features from the two modalities using two sub-modules, a patch alignment module, and cross-modality attention.

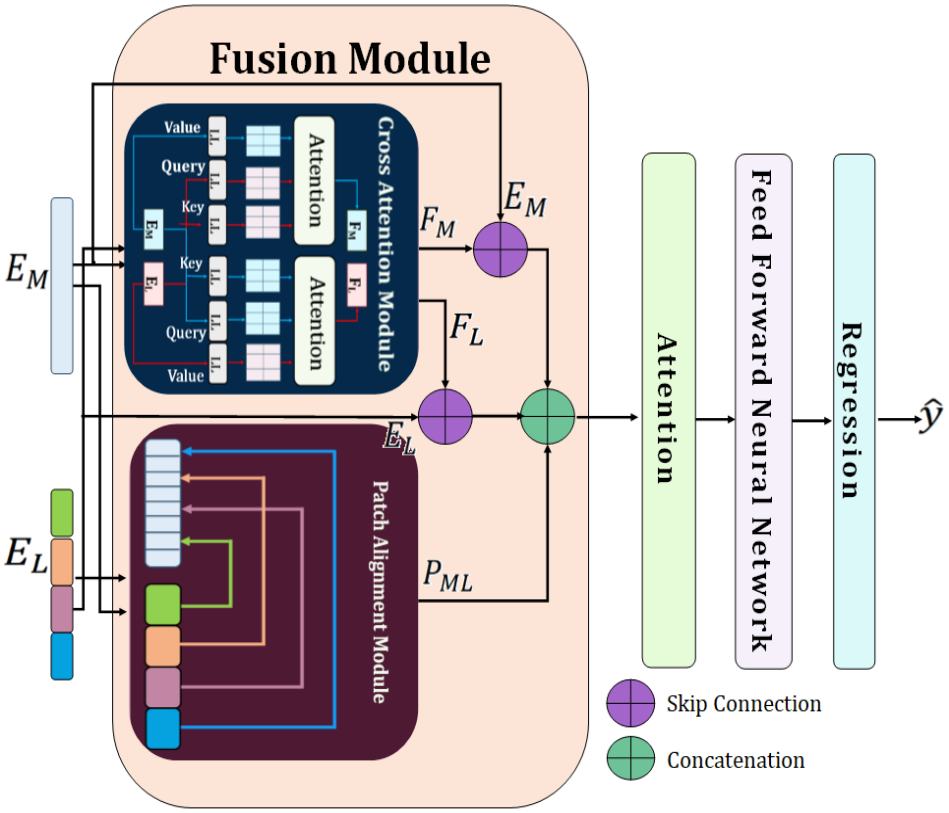

Figure 5.5: Fusion Module:FuSITSNet

Patch Alignment Module (PAM): We modified the Feature Alignment Module of LSFuseNet (see section 5.4.5) to work with image time series and named it as Patch

Alignment Module (PAM). PAM is used to align the patches of fine spatial resolution modality (Landsat-8) with the corresponding regions in the MODIS time series and learn fine temporal patterns for the aligned patches. This also suppresses the noise present in two SITS and mitigates its effect on the end task. In FAM the similarity calculation is bidirectional as spatiality is lost in both modalities. However, in PAM we calculate the similarity of Landsat patches with the MODIS image time series to know which patch aligns with which spatial region in the MODIS image time series. In the alignment process, we calculate the similarity, Sim_L of Landsat-8 patches with MODIS as given below:

$$Sim_L = E_L(E_M)^T \tag{5.7}$$

Softmax is applied over Sim_L and we use an average patch-wise aggregator over the MODIS embeddings as:

$$P^i_{ML} = softmax(Sim_{L_i})E_M, \ \ (1 \leqslant i < n) \tag{5.8}$$

$$P_{ML} = [P^1_{ML}; P^2_{ML}; \cdots ; P^n_{ML}] \tag{5.9}$$

n is number of patches traversed, P^i_{ML} is similarity-aware aggregated MODIS representation of i^{th} patch of Landsat.

Cross-Modal attention (CMA) learns the inter-modality relationships from the two embeddings E_M and E_L by applying bi-directional cross-modality attention by taking queries from both modalities to leverage their profound features. The scalar dot product attention between the hotspot spatial features of Landsat and highlighted temporal features of MODIS gives the joint high-quality features in both aspects. This helps the model to capture the complementary aspects of the two modalities and thus utilizes the information from one modality to compensate for the low quality of the other modality. Also, the module covers the untraversed Landsat-8 regions with the help of the MODIS time series. The output of the module is represented by F_M and F_L.

We concatenate P_{ML}, F_M, and F_L and apply multi-head self-attention to highlight the

combined hot spot features. It is followed by a feed-forward network comprising three linear layers with the Gaussian Error Linear Unit (GELU) activation function. Each linear layer is followed by layer normalization. Lastly, a regression layer is applied to get the prediction.

5.7 Learning Objectives

We have used two learning objectives for different modules in both models. The first objective is the marginal contrastive loss used in contrastive learning and the second is the mean square error (MSE) used for the downstream regression task.

5.7.1 Margin Contrastive Loss

Margin Contrastive Loss: In our twofold fusion technique, we innovatively applied margin contrastive loss [128]. Utilizing contrastive loss in regression problems is challenging since there are no explicit class categories to directly determine positive and negative pairs for training. The number of "classes" is roughly equivalent to the size of the dataset, rendering traditional contrastive loss implementation difficult. To overcome this challenge, we developed an approach to create positive and negative pairs based on locations and years, and then applied batch-wise margin contrastive loss. For each data instance (anchor) in a batch, we used the anchor as a positive instance by adjusting the dropout value in the encoder, while the negative pair is selected from a different location, a different year, or both.

The mathematical formulation of the loss is given below:

$$O_{anc} = FusionModel(X_i, dp_1) \tag{5.10}$$

$$O_{pos} = FusionModel(X_{j=i}, dp_2) \tag{5.11}$$

$$O_{neg} = FusionModel(X_{j\neq i}, dp_1) \tag{5.12}$$

O_{anc}, O_{pos}, and O_{neg} are the outputs of anchor, positive and negative instances, respectively. *FusionModel* is LSFuseNet or FuSITSNet. The loss is calculated as:

$$d_{ins} = \sqrt{(\sum_{i=1}^{g}(O_{anc} - O_{ins})^2 \ \forall \, all \ g \ instances \ in \ a \ batch} \tag{5.13}$$

$$loss_{ins} = \frac{1}{g}\sum_{i=1}^{g} max(min(d_{ins} \times target, margin), min) \tag{5.14}$$

$$where \ \ ins = \{pos, neg\}$$

5.7.2 Mean square error

The loss function used for regression is the mean squared error (MSE), which is a common metric to evaluate the difference between the predicted and actual values. The formula for MSE is shown below, where y_i represents the actual target, $\hat{y}_i$ represents the predicted output for a given location and year, and *N* represents the total number of location-year pairs considered.

$$MSE = \sum_{i=1}^{N} \frac{(y - \hat{y}_i)^2}{N} \tag{5.15}$$

Total Loss

The total loss (*L)* for the model is:

$$L = loss_{pos} + loss_{neg} + MSE \tag{5.16}$$

5.8 Models for Comparison

We have compared both models with different existing models. LSFuseNet is compared with the single modality and HLS data. FuSITSNet is compared with the single modality and existing generative fusion models. The details for both are given below:

5.8.1 Models for comparison: LSFuseNet

5.8.1.1 Baseline variants:

We tested a few variants of the proposed model for evaluating the importance of different modules used in the model. A brief description of the variants is given below:

MSTNet: We have used the simplified version of our proposed model for single modality and HLS data. The model variant consists of a surface reflectance band encoder $MSTE$, and the Task Specific Module (TSM). The variant is named MSTNet_L, MSTNet_S, and MSTNet_HLS for Landsat-8, Sentinel-2, and HLS data, respectively.

LSFuseNet_base: The baseline model of LSFuseNet consists of the surface reflectance band encoders $MSTE_L$, $MSTE_S$, and the fusion module (FM). The encoders are trained in an end-to-end training setting in our baseline.

LSFuseNet: The final version of the model includes the feature alignment module along with contrastive loss over the baseline and it uses the encoders, pre-trained on larger data which are fine-tuned during end-to-end learning along with other modules.

LSFuseNet+M: This model incorporates the additional meteorological data (if available) to further enhance the performance of the model.

LSFuseNet+M+S: The variant uses the additional soil data encoder (SDE) to incorporate available soil data along with the additional meteorological data.

5.8.1.2 Existing models for comparison:

We have compared our model with four existing models CNN [43], CNN+GP [43], and CNN+LSTM [44], and CYN [84]. The first two models use only surface reflectance data from a single satellite. The CNN+LSTM model uses soil data along with the surface

reflectance of a single satellite. These models are applied to different crops and locations in their original work. However, for a fair comparison, we have applied them for the same locations and time duration as that of our model. The hyperparameters of the models CNN, CNN+GP, and CNN+LSTM are fine-tuned using the Adam optimizer with learning rates of 0.001, 0.001, and 0.00001, respectively. While the CYN model uses SGD with a learning rate of 0.0001 and momentum of 0.7 for the optimization process.

5.8.2 Models for comparison: FuSITSNet

5.8.2.1 Baseline models:

To the best of our knowledge, there is no method that works with SITS for spatiotemporal problems. We applied the proposed PatchNet and TSE models on single modality image time series of Landsat-8 and MODIS, respectively, and compared them with FuSITSNet to see the significance of fusing two time series over a single modality.

5.8.2.2 Generative fusion models:

There are a few studies that fuse the information from one/two MODIS images with that of one/two Landsat images to generate a Landsat-like image at any timestamp. We generated images at every mid-timestamp to get time series of 8-day frequency using three such methods- STARFM [27], RSFN [118], and GAN[121]. We then applied PatchNet to get predictions using enhanced SITS and compared them with the proposed FuSITSNet model.

STARFM is a pixel-based method that works on the principle of a moving window and calculates the value of the pixel depending on the value of neighboring pixels. The method assumes that if MODIS and Landsat surface reflectances are equal at a given time, then these values should be equal at the prediction date. Predicting the value for each pixel one by one is a very time-consuming task and it took us almost 25 days on an A30 GPU server with 24GB RAM to generate the data at mid-timestamps of the Landsat-8 time series for all the counties used in crop yield prediction. As the process was very inefficient and was not the best performer in the generative models, so we did not generate the data for counties used in snow cover prediction.

Robust Spatiotemporal Fusion Network (RSFN) is a generative adversarial network and attention mechanism-based model to generate the Landsat-like image using one coarse-resolution (MODIS) image on the prediction date and two referential fine-resolution images (Landsat-8) before and after the prediction date as model inputs. It took us around 7 days to train the model on an A30 GPU server with 24GB RAM and 5 days to generate the data for all the counties used in the three applications.

Conditional Generative Adversarial Network is used to generate cloud-free Landsat-like images using MODIS image at time *t* and time *t*-1 and Landsat-8 image from time *t*-1 to predict the image at time *t*. A U-Net-based generator is used in the model to fuse the two images. It took us approximately 5 days to train the model on an A30 GPU server with 24GB RAM and 4 days to generate the data at mid-timestamps of the Landsat-8 time series for all the counties used in the three applications.

The generated sample images by the three models are given in Figure 5.6.

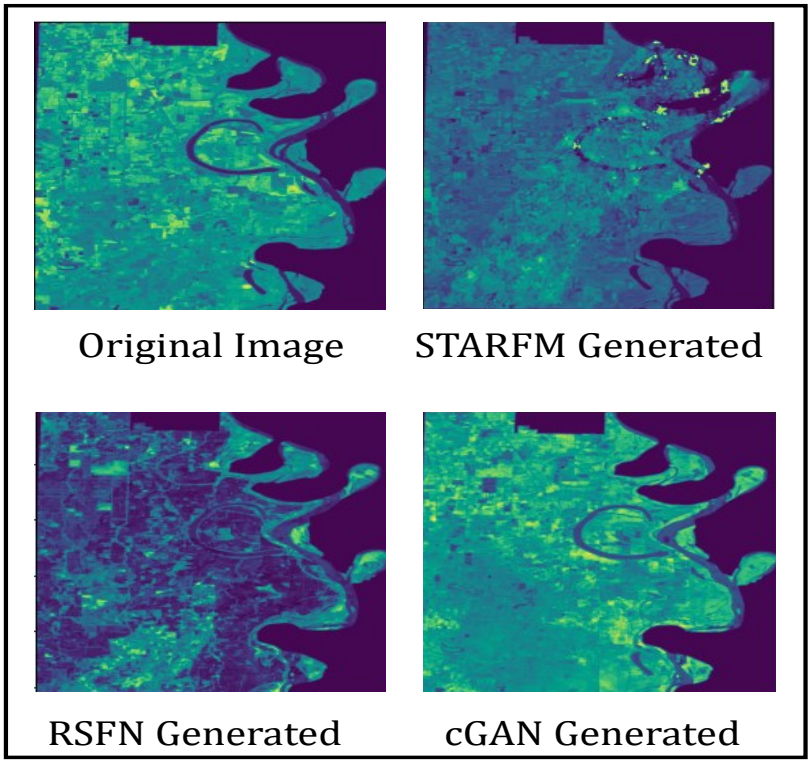

Figure 5.6: Original Landsat-8 image and images generated by different generative models

5.9 Experimental Setup

We performed experiments using Pytorch 1.11.0 and CUDA 11.7 on an A100 GPU server with 80 GB RAM for all the models. The optimizers used for both models are different.

While training LSFuseNet for CYP, the stochastic gradient descent (SGD) optimizer was used with a learning rate of 0.0001 and a momentum of 0.7. The model was trained for 150 epochs with a batch size of 16, while for SCP, the same optimizer was used with a learning rate of 0.00001 and the same momentum value. The model was trained for 50 epochs with a batch size of 64.

FuSITSNet is trained for 50 epochs with a batch size of 8 using Adam optimizer with a learning rate η. We have trained the model with 5 years of data (2014-2018) and, 2 years (2019 and 2020) for testing. To predict the output for the z^{th} year, the training is conducted until the $(z-1)^{th}$ year. For CYP, $\eta = 0.0005$ for a single modality (TSE and PatchNet) and $\eta = 0.000005$ for FuSITSNet. In case of SCP and SEP, $\eta = 0.00001$ for all three models. The evaluation metric used is Root Mean Squared Error (RMSE).

5.9.1 Evaluation Metrics:

Two metrics root mean square error (RMSE) and mean absolute error (MAE) are used for the performance evaluation of the proposed model for both applications. MAE represents the mean magnitude of the errors that occurred in the predictions by the model, while RMSE measures the root mean squared magnitude of the errors, giving more weight to larger errors. A higher RMSE indicates that the model is suffering from by a substantial quantity of errors with significant magnitudes

The higher RMSE depicts that the model suffers from a large number of large-magnitude errors. The crop yield is measured in bushels/acre (bu/ac) and snow cover is measured as the percentage of area under snow.

The formula for RMSE is given as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_{c,z} - \hat{y}_{c,z}^{i})^2}{N}} \tag{5.17}$$

where *N* is the total number of $\{c, z\}$ pairs, *c* is the county, *z* is year, month, or fortnight for CYP, SCP, or SEP, respectively.

5.10 Results and Discussion

In this section, we present results and discussion for both models in subsequent subsections.

5.10.1 Results: LSFuseNet

The first set of experiments is conducted to compare the predictions of MSTNet (using Landsat-8, Sentinel-2, and HLS data) and LSFuseNet. Table 5.2 depicts MAE and RMSE for both applications (CYP and SCP). We observed that Sentinel-2 performed slightly better than Landsat-8 when single satellite data is used individually. This is because of its finer spatial and temporal resolution than Landsat-8. We also compare the results using the existing HLS data which is also a fused data of Landsat-8 and Sentinel-2. The model performs better for HLS than individual Landsat-8 and Sentinel-2 for both applications due to a finer temporal resolution of HLS in comparison to Landsat-8 and Sentinel-2. The HLS data shows that RMSE is improved as compared to Sentinel-2 by 2.5% in the case of SCP and the improvement of 2.3% and 0.32% is observed for soybean and corn yield prediction, respectively.

For CYP, MAE achieved by LSFuseNet is 18.6927 and 6.1953 for corn and soybean, respectively and the respective RMSE is 18.4671 and 5.9677 bu/ac. The MAE is approximately 23%, 21%, and 13% improved from that of Landsat-8, Sentinel-2, and HLS data, respectively for soybean yield prediction. It can be observed that LSFuseNet improves RMSE by 23.5% and 22.9% from HLS for soybean and corn, respectively.

A similar pattern is observed in SCP where the achieved MAE is 10.6437 which

Table 5.2: Comparison of MAE and RMSE using single satellite data and LSFuseNet

Model	**CYP**				**SCP**	
	Corn		Soybean			
	MAE	RMSE	MAE	RMSE	MAE	RMSE
MSTNet_L	19.673	24.064	8.056	8.043	13.257	19.478
MSTNet_S	19.353	24.033	7.939	7.984	12.748	19.269
MSTNet_HLS	19.318	23.954	7.201	7.802	12.679	18.769
LSFuseNet	**18.693**	**18.467**	**6.195**	**5.968**	**10.644**	**10.322**

Table 5.3: Difference between ground truth and predicted yield for corn (bu/ac)

S. No.	**Location**	**Gound Truth**	**MSTNet_L**	**MSTNet_S**	**MSTNet_HLS**	**LSFuseNet**
1	17-1	178.4	8.602	-6.707	4.557	**-0.412**
2	17-105	182.4	8.293	4.002	2.933	**-0.988**
3	5–79	165.0	-2.681	-2.739	1.256	**0.982**
4	27-91	204.5	162.946	123.797	121.216	**19.250**
5	28-3	153.5	101.946	89.875	74.35	**-11.343**

shows 19%, 16.5%, and 16% improvement on Landsat-8, Sentinel-2, and HLS data predictions, respectively. The improvement in RMSE is 46.43% and 45% as compared to the predictions of Sentinel-2 and HLS data, respectively.

The increased rate of improvement in RMSE in comparison to MAE in the proposed model shows that LSFuseNet is capable of better capturing the variance in the data. Also, we can observe that the RMSE of LSFuseNet is less than the respective MAE in all three cases. The reason behind it is the reduction in the large errors more than that of small errors. For example, as shown in Table 5.3, the error for locations 1-3 is less in magnitude for all the models with the least in LSFuseNet. Whereas, the error in locations 4 and 5, is quite high for single modality and HLS models which are drastically reduced for LSFuseNet. For example, the least error 121.216 of HLS is reduced to 19.250 for the county '27-91', and from 74.34 of HLS in the county '28-3' is reduced to 11.343.

It can be observed from Table 5.2 that the reduction in RMSE obtained by LSFuseNet from that of a single modality for SCP is much higher than that of CYP. This can be explained that in CYP, data unavailability at finer temporal granularity does not affect

Table 5.4: Ablation Study: LSFuseNet (RMSE for CYP and SCP)

Model	CYP		SCP
	Corn	Soybean	
LSFuseNet_base	19.7568	7.5602	13.0453
LSFuseNet_base+FAM	18.7399	6.2649	10.9946
LSFuseNet_base+Pretrain	18.8718	6.6974	11.4332
LSFuseNet	18.4671	5.9677	10.3218

Table 5.5: RMSE obtained by LSFuseNet for different pretraining tasks

Pretraining Task	**Corn**	**Soybean**	**Snowcover**
Without Pre-training	18.7399	6.2649	10.9946
Reverse TS	18.5584	5.9977	10.8641
Irregular TS	18.5496	5.9847	10.7794
Reverse & Irregular TS	18.4671	5.9677	10.3218

more as it can be covered by analyzing the crop health in the upcoming timestamps. Whereas in SCP this effect is much more. It is clear that this impact is reduced by the fusion of two modalities via LSFuseNet.

Ablation Study: Table 5.4 presents the results obtained without various modules of the proposed model. It can be observed that the RMSE for SCP using the baseline model is 13.0453 which is reduced to 10.9946 ($\approx$ 15%) after using the FAM. Similarly, the RMSE is reduced to 11.4332 ($\approx$ 12%) when we pre-trained our baseline model. When we pre-trained the model using FAM (LSFuseNet), collectively, they reduced the RMSE to 10.3218 which is $\approx$ 21%. The results show the importance of enhancing the features of each modality by guiding them from the other modality and mitigating the effect of any noise using FAM. Similar results were obtained for CYP.

Table 5.5 presents the results obtained with different pre-training tasks. The best results were obtained when the encoders were pre-trained using both classification tasks (see section 5.5). Also, it is evident from the results that pre-training has substantially improved the RMSE by 12% after pertaining the baseline for SCP and the improvement

observed after pertaining the model using FAM is 6%. Similar patterns are observed for CYP.

Importance of Feature Level Fusion: We also experimented with fusing only those bands which are common in both satellites. To carry out these results we have used the non-pre-trained version of our model. The RMSE obtained using only the common bands in CYP is 7.0535 and 19.6072 bu/ac for soybean and corn, respectively, and for SCP it is 12.7935. On comparing these RMSE values with that of the fusion of two modalities with all bands, the performance degraded by approximately 14.06% for SCP and 11.18% and 4.4% for corn and soybean yield prediction, respectively. This observation indicates the importance of using all the bands in the satellites. This kind of fusion is not possible in the case of pixel-level fusion. Thus, these results clearly emphasize the importance of a feature-level fusion technique to maximize the contribution of the spectral resolution of the satellites

Zero Shot Testing: We also experimented with zero-shot learning in which testing is performed on the data from the classes which are completely different from that of testing classes. For CYP, we tested the model for 79 counties after training it for a different set of 500 counties. The RMSE obtained by LSFuseNet is 18.7165 and 6.0544 bu/ac for corn and soybean, respectively. Similarly, in the case of snow cover, we trained the model for 1000 counties, tested it for 312 counties, and achieved the RMSE of 10.6088. The results obtained by LSFuseNet in zero-shot learning are comparable to that obtained during normal training (18.4671 and 5.9677 for corn and soybean yield prediction and 10.3218 for SCP).

Incorporating different modalities: We modified LSFuseNet to incorporate meteorological and soil data along with surface reflectance data. The results given in Figure 5.7 show the impact of using different modalities in different models for both applications. It can be observed from the results that RMSE improves when meteorological and soil data is incorporated into both single satellite models and fusion models. There is an improvement of $\approx$ 2%, 4%, 7%, and 15% in corn yield prediction when meteorological data is incorporated into MSTNet_L, MSTNet_S, MSTNet_HLS, and LSFuseNet, respectively.

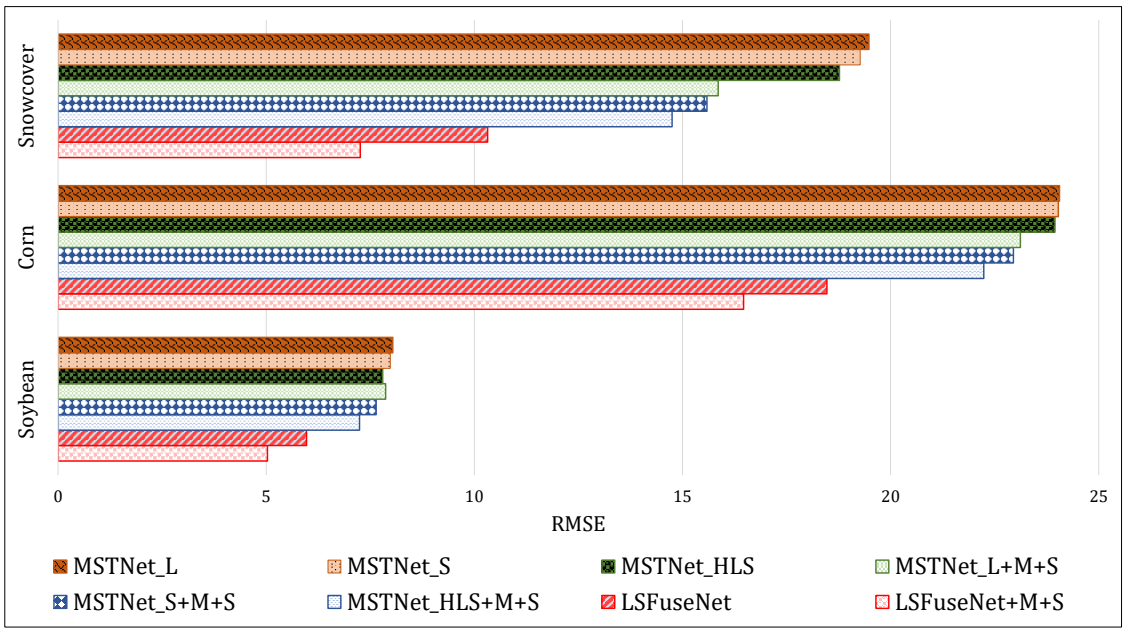

Figure 5.7: Incorporating additional modalities-effect on RMSE.
*LSFuseNet+M+S represents only using meteorological data in case of Snowcover Prediction

Table 5.6: Comparison of LSFuseNet with existing Models for CYP (in RMSE)

Model	Corn	Soybean
CNN [43]	26.997	9.679
CNN + GP [43]	31.049	8.383
CNN + LSTM [44]	25.477	10.956
CYN [84]	18.938	5.990
LSFuseNet	**18.467**	**5.967**
LSFuseNet+M+S	**16.467**	**5.029**

It is evident from the results that maximum improvement is obtained using our proposed model. Similar results were observed for soybean yield prediction and SCP.

Comparison with existing models: In Table 5.6, we have compared the results with the existing models for crop yield prediction. All these existing models are designed to work with data from a single satellite. The results in the table for these models are taken for Sentinel-2. CNN and CNN + GP models are designed to predict the yield using only the surface reflectance data from a satellite. So, for a fair comparison, we compared it with LSFuseNet and observed a significant difference in the RMSE. The RMSE achieved

by CNN and CNN+GP is 26.997 and 31.049, respectively for corn and in the case of soybean, it is 9.679 and 8.383 bu/ac respectively. The corresponding RMSE obtained by LSFuseNet is 18.467 and 5.968 bu/ac for corn and soybean, respectively. Our results are even better than the other two models, CNN+LSTM and CYN which use additional meteorological and soil data. If we compare these results with LSFuseNet+M+S, the error is reduced to 16.467 and 5.029 bu/ac for corn and soybean, respectively.

We have also compared the results of CYN using MODIS data for 16 years [84]. The RMSE of CYN with MODIS data is 21.505 bu/ac and 8.166 bu/ac for corn and soybean, respectively which is much much larger than that of LSFuseNet+M+S.

5.10.2 Results: FuSITSNet

Comparison of FuSITSNet with single modality baselines: Table 5.7 presents RMSE obtained by FuSITSNet and single modality baselines TSE and PatchNet using MODIS and Landsat-8 time series, respectively. It is evident from the results, that the PatchNet (Landsat-8) performed better than TSE (MODIS) with $\approx 8\%$ and 3.5% lower RMSE in corn and soybean yield prediction, respectively. RMSE reduced from 17.17 to 12.81 for SCP and from 8.86 to 8.54 for SEP, making an improvement of 25% and 3.6%, respectively. This shows the importance of using high spatial resolution data for the applications. The results improved further using FuSITSNet for all three applications. RMSE reduced by $\approx 24\%$ and 30% for corn and soybean yield prediction, respectively when compared with PatchNet (Landsat-8). A similar pattern is observed in SCP with an improvement of 46% from TSE and $\approx 28\%$ from PatchNet (Landsat-8). The maximum improvement is observed in SEP with almost 76%. The huge reduction in error signifies that FuSITSNet exploits high temporal and high spatial features and is thus suitable for spatiotemporal applications.

Our baselines on enhanced SITS: We generated Landsat-8 images at every mid-timestamp. It can be observed from Table 5.7 that RMSE reduces when PatchNet is applied over enhanced time series in comparison to PatchNet (Landsat-8) with an exception in the case of corn yield prediction. Out of three generative models, GAN generated

Table 5.7: FuSITSNet vs single modality baselines

Model	CYP		SCP	SEP
	Corn	Soybean		
TSE (MODIS)	23.3352	7.5457	17.1675	8.8633
PatchNet(Landsat-8)	21.4697	7.2901	12.8136	8.5436
PatchNet(STARFM) [27]	20.2898	6.3084	—	7.2274
PatchNet(RSFN) [118]	22.8389	6.4323	12.3298	8.0126
PatchNet(GAN)[121]	18.102	6.296	11.9518	7.0431
FuSITSNet	**16.1925**	**5.0389**	**9.2308**	**2.0447**

image time series performed the best with an improvement of ≈16%, 14%, 7%, and 18% for corn, soybean, snow cover, and solar energy prediction, respectively.

Comparison of FuSITSNet with Generative fusion models: Table 5.7 shows that FuSITSNet outperforms all the scenarios where PatchNet applied on SITS generated by the existing generative fusion models. RMSE reduced by $\approx 20\%$ for CYP in comparison to the pixel-based model STARFM. The reduction in RMSE for FuSITSNet is 29% and 10% in comparison to learning-based models RSFN and GAN, respectively for corn yield prediction and the corresponding reduction is 21% and 19% for soybean. For the snow cover application, FuSITSNet performed 23% better than PatchNet(GAN), the top performer. The best results are obtained for solar energy prediction where RMSE is reduced by $\approx 70\%$ using FuSITSNet than that of PatchNet(GAN).

Comparison of running time and the number of parameters: We have compared the no. of parameters & running time required for each model (see table 5.8). In PatchNet(RSFN) and PatchNet(GAN), the proposed model PatchNet is applied to images generated by existing RSFN (Tan et al. 2022), and GAN (Bouabid et al. 2020) generative models. Thus, the no. of parameters & training time is the sum of parameters & time needed in the generation and prediction process. FuSITSNet has more parameters, but the running time is approx. $1/4^{th}$ of other fusion models as it does not need a generation process. Moreover, it outperforms the prediction. The inference time of each model is almost the same which is ≈ 10 mins.

Table 5.8: Running time & No. of training parameters for models for CYP: corn

	Parameters	Training time (hrs)	RMSE
TSE (Modis)	183M	30	23.3352
PatchNet (Landsat-8)	131M	45.5	21.4697
PatchNet (RSFN)	146M (generation)+ 205M (prediction) = 351M (total)	168 (RSFN training)+ 120 (generation)+ 50 (prediction) = 338 (total)	22.8389
PatchNet (GAN)	165M (generation)+ 205M (prediction) = 370M (total)	120 (GAN training)+ 96 (generation)+ 50 (prediction) = 266 (total)	18.102
FuSITSNet	416M	70	16.1925

Table 5.9: Ablation Study: FuSITSNet

Model	**CYP**		**SCP**	**SEP**
	Corn	**Soybean**		
FuSITSNet	16.1925	5.0389	9.2308	2.0447
FuSITSNet (no PAM)	17.1989	5.2862	12.4484	2.5017
FuSITSNet (no CMA)	17.2425	5.8476	11.5232	2.7494

Ablation Study: We carried out an ablation study to show the importance of modules in FuSITSNet. Considering variations in results in Table 5.9, we observed that RMSE increased significantly without using PAM with a maximum increase of 25% for SCP followed by 18% in solar energy. Similarly, the performance of the model is also degraded without cross-modality attention. This shows that both modules are important to effectively exploit high spatial and high temporal features in the two SITS.

Hyper-parameter Tuning: We experimented with η ranging from 0.0001 to 0.007 for TSE & PatchNet & $\eta = 0.0001\ to\ 0.000005$ for FuSITSNet for CYP. For SEP and SCP, we experimented for $\eta = [0.0001\ to\ 0.00003]$ for all three models. Final values were decided based on validation & training graphs.

We experimented with 1000, 500, 400 & 100 hidden layers (HL). HL is fixed to 100 based on the trade-off between computation time and RMSE. GPU memory overflowed when HL exceeded 500. We experimented with different kernel sizes for both TSE and

Table 5.10: Incorporating Meteorological attributes

Model	Without meteorological	With meteorological
Corn		
TSE (MODIS)	23.3352	22.7405
PatchNET	21.4697	20.6316
FuSITSNet	16.1925	15.8448
Soybean		
TSE (MODIS)	7.5457	7.4957
PatchNET	7.2901	6.9631
FuSITSNet	5.0389	4.9197

PatchNet. For PatchNet the experiments were conducted with kernel sizes starting with timestamp length 5 and reduced to 3 and $H \times W$ from 11 to 7. Best results were obtained for $3 \times 9 \times 9$. Similarly, kernel size for TSE is fixed to $5 \times 23 \times 39$.

Significance of Meteorological data: Meteorological conditions have a significant impact on the spatiotemporal applications considered in this thesis. As in the case of crop yield prediction, weather conditions in the area throughout the crop cycle can have a positive or negative impact on the crop yield. Therefore, we incorporated meteorological data into PatchNeT and FuSITSNET to see its impact on crop yield prediction. We used an additional LSTM layer to learn the patterns from the meteorological attributes and then concatenated the LSTM output with the image time series encoder output to get the predictions using feed forward and regression modules in the two models. From the results, it is evident that the model performance improved when meteorological data is incorporated. The results are given in Table 5.10.

5.11 Main Contributions

In this chapter, we have given two fusion techniques LSFuseNet and FuSITSNet to handle the spatial and temporal resolution trade-offs faced by satellite systems. LSFuseNet works with histogram time series and FuSITSNet works with image time series obtained from two different satellite systems. The main contributions of the chapter are:

- We proposed two techniques which perform a feature-level fusion of multi-spectral

varying temporal and spatial resolution satellite time series data.

- LSFuseNet is a dual-fusion technique of fusing histograms using two key modules - a fusion module and a novel feature alignment module which together have the capability of learning enhanced fine-grained features and mitigating the noise.

- A pre-trained encoder of LSFuseNet on satellite data for two classification tasks to enhance the model performance. To the best of our knowledge, this is the first attempt at pre-training in the field of satellite data.

- FuSITSNet is a twofold feature-based fusion model for fusing two image time series having different resolutions. We complementarily use a patch alignment module and cross-modality attention to learn high spatial resolution features of Landsat-8 and high temporal features of MODIS.

The proposed fusion techniques are capable of achieving the following: 1) working with multi-spectral high-resolution satellite images without requiring high-end hardware; 2) we can use multi-spectral time series of varying temporal, spatial, and spectral resolutions; 3) achieving better results in comparison to the single satellite data; 4) The direct feature-based learning from two SITS using proposed techniques outperform the enhanced SITS obtained from existing generative fusion models.

5.12 Summary

The democratization of satellite imaging technology is still marred by the need for processing huge volumes of data and by the unavailability of high spatial and temporal resolution images from a single publicly available satellite system. We proposed two fusion techniques to handle the trade-off between spatial and temporal resolution of satellite systems. LSFuseNet and FuSITSNet fuse two image time series to obtain a joint representation that captures the high spatial resolution features of one satellite system and the high temporal resolution features of another. The difference is that LSFuseNet works with histogram time series and FuSITSNet works with image time series. Both the techniques

are twofold fusion techniques having two key modules-fusion module (FM) and a feature alignment module (FAM) in LSFuseNet and Cross-Modal attention (CMA) and Patch Alignment Module (PAM) in FuSITSNet. The two modules reinforce each other in their respective models. Together these modules, highlight and exchange the hotspot information in two modalities, learn fine-grained features, and mitigate noise in the data. The highlight of the fusion techniques is that high spatial and temporal features are learned without image generation, thereby not increasing the voluminous data further as is the case with generative approaches.

Chapter 6

SpInN: A broader perspective for Spectral Reflectance Indices

[1]

6.1 Introduction

Spectral Reflectance Indices (SRIs) designed by physicists and domain experts, are mathematical formulations of bands in the visible and near-infrared electromagnetic spectrum. The formulations of these bands improve the sensitivity towards the detection of vegetation, environmental variables, physiological and morphological characteristics of the earth's surface or plants, etc. [45]. Use of satellite-obtained SRIs is a ubiquitous practice in monitoring vegetation health [39, 85–87].For example, the Normalized Difference Vegetation Index (NDVI), derived from Red and Near Infra-red (NIR) bands has emerged as one of the most commonly employed spectral indices for monitoring crop growth and crop yield. However, various other SRIs have been used to support crop yield estimations like the Normalized Difference Water Index (NDWI), Enhanced Vegetation Index (EVI),

[1] *The work presented in this chapter is communicated:*

- *Arshveer Kaur, Poonam Goyal, Vansh Bansal, Deep Pandya, and Navneet Goyal, "SpInN: A Pre-trained Model for Spectral Indices Recommendation for Earth Observation Applications using Satellite Data", in IEEE Transactions on Neural Network and Learning Systems. [Communicated]*

etc. [46]. Different SRIs highlight different characteristics depending upon their mathematical formulations and thus can be useful in different applications. However, this has not been explored yet. We bolstered the idea of using SRIs for solving applications like snow cover, soil moisture prediction, land cover classification, etc.

Identification of relevant SRIs and their usefulness for an application needs domain expertise. However, to the best of our knowledge, there are no automated ways to recommend relevant SRIs for a given application. Existing studies aggregate SRIs values for a location (like a county) at a given timestamp resulting in a univariate time series for a location, thereby losing critical spatial information. Some researchers have used multiple SRIs without examining their relevance to the application, and their compatibility with each other leading to noise and eventually degraded performance.

To handle the problem of loss of spatial information, we propose to calculate the spectral index at every pixel of the satellite image as per the spatial resolution of the satellite, thereby, generating a matrix of spectral reflectance index values for a location at a timestamp. These matrices are stacked to obtain an SRI image having the number of channels equal to the number of SRIs considered. We propose a model **Sp**ectral **In**dex **N**etwork (SpInN), which selects the most relevant spectral indices for a given application.

We tested SpInN on six earth observation applications viz. crop yield prediction (for two crops corn and soybean), soil moisture, solar energy, snow cover, cloud cover prediction, and land cover classification. The results indicate that recommendations made by the proposed model produce state-of-the-art predictions.

6.2 Related Work

Satellite image technology is witnessing significant improvements in its ability to capture data at high spatial, temporal, and spectral resolutions. This facilitates monitoring of the Earth's surface at desired levels of resolution for different Earth observation applications. Many of the satellite systems provide data free of cost e.g. MODIS [10], Landsat-8/9 [40], Sentinel [41], etc. Satellite imagery is a cost-effective technology to get data at a global scale. Most of the applications require time series analytics. The researchers

have used satellite data in the form of histograms [43, 44, 58], spectral reflectance indices [32, 33] or images [18] for different applications. We focused on six earth observation applications viz. prediction of crop yield, solar energy, soil moisture, cloud cover, snow cover, and land lover classification.

Crop Yield Prediction: Timely and accurate yield estimation helps the government to make various decisions regarding insurance, import-export, etc. Different studies have applied various statistical [129] and machine learning [46, 130–133] on different types of data. For example, systems [59, 61] predicted crop yield using meteorological data, in [43, 44, 58, 77] used histogram time series of satellite data. Authors have used a few common spectral indices including NDVI, EVI, NDWI, SAVI, etc. for CYP [131, 134]. All these studies have taken the spectral indices as a single value by averaging the entire image for a region at a timestamp. One of the significant drawbacks is the loss of spatial information, as this approach discards the arrangement of pixels and the spatial relationships within the image. Moreover, noise, outliers, or irrelevant details may affect the global mean rather than remain localized.

Snow Cover Prediction: Traditionally, meteorological stations measured snow depth through manual surveys. In recent years, researchers have used time series data obtained from satellites MODIS, SPOT-4/5, and Landsat-8 using fractional and binary snow-cover data [135, 136]. MODIS no longer provides this data and thus recent studies [96, 137] used normalized difference snow index (NDSI) to predict snow cover.

Solar Energy Prediction: Predicting solar energy is essential for identifying optimal locations for solar plant installations, aiming to decrease reliance on fossil fuels and foster economic development. Researchers have explored the prediction of solar energy using meteorological attributes in recent studies [98, 101–103] using different machine learning (RF and SVM) and deep learning (LSTM, ANN, DNN) models. For example, a study [103] is performed over Basque country using ANN applied on the weather data and achieved RMSE ranging from 5.33 to 77.76 on different days. In [98], SVM and DNN are applied over meteorological data of Turkish provinces, and the maximum RMSE is observed as 2.820 and 2.814, respectively.

Cloud Cover Prediction: Accurate forecasting of cloud cover is critical for various applications, like photovoltaic energy production, agriculture, tourism, etc. A study [138] has done cloud cover prediction using meteorological data using regression models, and ground-based sky images [139, 140] using CNN models. Ground-based cameras are a costly method and have limitations, such as fixed positions, inconsistent data quality, calibration problems, etc. To overcome these issues, researchers [137, 141] used satellite data and deep learning models like DNN and CNN with regression models.

Soil Moisture Prediction: Accurate prediction of soil moisture [142] is crucial for efficient water resource management in agriculture and hydrological cycles. Researchers [142] conducted the study in a small region of Beijing for soil moisture prediction using meteorological data and achieved the RMSE of 4.05, and 6.01 using SVM, and ANN, respectively. FLUXNET dataset incorporating meteorological and soil parameters has been used in [143] and [144] to predict soil moisture by applying LSTM. Taktikou et al. [145] used NDVI combined with sensor and land surface temperature to retrieve daily soil moisture content using a regression model.

Land Cover Classification: It is a current research interest as it plays an important role in land use analysis, urban planning, etc. The problem has been attempted [15–19] using satellite images and with ground truth pixel-wise annotated images given in datasets MCD12Q1 V6, GlobCover2009, GLC, and GlobeLand30. These datasets have different classes and have different spatial resolutions of the annotated images. However, data is given at the yearly temporal granularity. For example, authors achieved an accuracy of 82.7% [18] for Europe using ground truth divided into 5 classes. Some other studies working with the MCD12Q1 dataset have obtained accuracy 74.8% [146] and 69% [146] respectively for regions in China. In another study [20] conducted in Northeast Asia, researchers obtained an accuracy of 77.94% for the MCD12Q1 data classified into 17 classes.

From the literature survey, it is clear that the models used for applications other than CYP are either shallow machine learning models or simply neural networks which are not capable of modeling complex patterns of satellite data. There is no generic model

which can be seamlessly applied across applications with desired output. The proposed model SpInN aims to overcome the aforementioned limitations.

6.3 Study Area and Data

6.3.1 Study Area

This study encompasses a comprehensive geographic scope, spanning the counties of the United States. We used different counties for different applications. The focus for Crop Yield Prediction (CYP) is set upon those counties reigning as major producers of corn and soybean. This elite cohort is drawn from states of considerable agricultural prominence, including but not limited to states viz. Illinois, Tennessee, Iowa, Missouri, Nebraska, Michigan, Wisconsin, Mississippi, etc. In the realm of Snow Cover Prediction (SCP), the focus shifts to counties that witness an average snowfall of more than 250 inches. The counties lie in the states of Washington, Oregon, Utah, California, New Hemisphere, and Colorado. In the context of applications like Solar Energy Prediction, Soil Moisture Prediction, Cloud Cover Prediction, and land cover classification, we selected a quintet of states comprising Illinois, Indiana, Iowa, Kansas, and Kentucky. These states take center stage in these predictive endeavors, representing regions where the interplay of environmental factors holds particular significance for the applications under consideration.

6.3.2 Data Used

The input data used consists of spectral reflectance index images, and additional meteorological data, and soil data. The ground truth data is specific to the application. A brief description of each type of data is given below:

6.3.2.1 SRI images:

Spectral reflectance indices represent a diverse array of mathematical metrics, ratios, or linear combinations. We have derived all the spectral indices used from MODIS product MOD06 which has a spatial resolution of 500m and a revisit time of 8 days. We have used spectral indices given in Table 6.1. Other details are given in Chapter 2 section 2.5.3.

Table 6.1: List of Spectral Reflectance Indices considered

Spectral Index	Formula	Key Trait
NDVI	$\frac{(NIR-Red)}{(NIR+Red)}$	Vegetation health
NDWI	$\frac{(Green-NIR)}{(Green+NIR)}$	differentiate water from dry land
SAVI	$\frac{(NIR-Red)}{(NIR+Red+L)} * (1+L), L\ is\ constant$	dense vegetation
NDYI	$\frac{(Green-Blue)}{(Green+Blue)}$	overcame limitations of NDVI
PSRI	$\frac{(Red-Green)}{NIR}$	vegetation health monitoring
EVI	$G \times \frac{(NIR-Red)}{(NIR+C1\times Red-C2\times Blue+L)}$ where G, C1, C2, L are constants	quantify vegetation greenness
SR	$\frac{NIR}{Red}$	differentiate bare soil and vegetation
WDRVI	$\frac{(a\times NIR-Red)}{(a\times NIR+Red)}$, $0.1 \leqslant a \leqslant 0.2$	moderate-to-high vegetation density when NDVI exceeds 0.6.
MSAVI	$\frac{2*NIR+1-\sqrt{(2*NIR+1)^2-(8*(NIR-Red))}}{2}$	used as substitute NDVI where it fails due to low vegetation or a lack of chlorophyll in the plants
MSR	$\frac{\frac{NIR}{Red}-1}{\sqrt{\frac{NIR}{Red}}+1}$	produces images with good contrast and a high signal-to-noise ratio

6.3.2.2 Ground Truths:

The crop yield data for U.S. counties is taken from Quick Stats [80]. For snow cover, normalized difference snow index (NDSI) has been acquired at the county level using the MODIS product MOD10A1 [104]. The soil moisture data is acquired at the county level every month using NASA-USDA Enhanced SMAP Global soil moisture data [147]. The information about solar energy produced in a county has been acquired from weather data[105]. Daily cloud cover values are typically derived by calculating the mean of the hourly cloud coverage values throughout the day. We acquired data captured on a daily basis from [105]. For land cover classification, we use data from MCD12Q1 [148] which classifies land cover into 17 classes, assigning a class label to every pixel of the satellite image. The collection of ground truths for every application is given below:

Yield data: The crop yield data for U.S. counties utilized in this study is taken from Quick Stats [80], a comprehensive database compiled by the United States Department of Agriculture (USDA). The data used in this study spans the period from 2002 to 2020. The yield values are quantified in bushels per acre (bu/ac), offering a standardized unit for assessing and comparing crop productivity across different regions.

Snow cover data: The snow cover data has been acquired at the county level using the MODIS product MOD10A1 [104]. This dataset offers information about snow cover extent with a spatial resolution of 500 meters. The methodology employed in this process relies on the normalized difference snow index (NDSI). It is a metric that exploits the differences in reflectance values between surfaces covered by snow and those without snow in the visible and near-infrared spectral regions. NDSI is a key indicator, representing the percentage of the area covered by snow.

Soil Moisture Data: The soil moisture data is acquired at the county level for every month using NASA-USDA Enhanced SMAP Global soil moisture data [147]. This dataset was developed by the Hydrological Science Laboratory at NASA's Goddard Space Flight Center in cooperation with USDA Foreign Agricultural Services and USDA Hydrology and Remote Sensing Lab. The Soil Moisture Active Passive (SMAP) instrument measures the amount of water in the surface soil everywhere on Earth. The value repre-

sents the amount of water in mm. The data is available from 2016 and thus data used in this study spans the period from 2016 to 2020.

Solar energy data: The information about solar energy produced in a county has been acquired from [105]. The value represents the total solar energy produced in MJ/m2 for a county in a day. The data used in this study spans the period from 2014 to 2020.

Cloud cover data: Cloud cover represents the extent of the sky covered by clouds and is usually expressed as a percentage. This percentage encompasses clouds at all altitudes. Daily cloud cover values are typically derived by calculating the mean of the hourly cloud coverage values throughout the day. Essentially, it represents the proportion of the sky covered by clouds throughout the day. We acquired data captured on a daily basis from [105]. The data used in this study spans the period from 2014 to 2020.

Land cover classification: Land cover can be classified into a different number of classes. We used MCD12Q1 [148] which classifies land cover into 17 classes. The dataset provides annotations for the land cover at a spatial resolution of 500m and yearly granularity. The data used in this study spans the period from 2014 to 2020.

6.3.3 Data Preparation

We used Google Earth Engine to download satellite images and performed basic pre-processing steps given in Chapter Section 2.4. Meteorological data is downloaded and pre-processed using the steps given in Chapter 2 section 2.6.

Deciding Time series length for each application: The length of the time series varies for each application. For crop yield prediction, a padded crop cycle is employed for each crop, extending two timestamps on either side (Chapter 3). For the other applications viz. snow cover, soil moisture, solar energy, and cloud cover prediction, the time series spans the last three months to forecast the target. For snow cover and soil moisture, prediction is made for the upcoming month, and for solar energy and cloud cover, prediction is done for the subsequent fortnight. These applications are sensitive to climatic conditions that undergo changes over short durations. Thus, the length of the time series is small. For land cover classification, prediction is done at the yearly granularity.

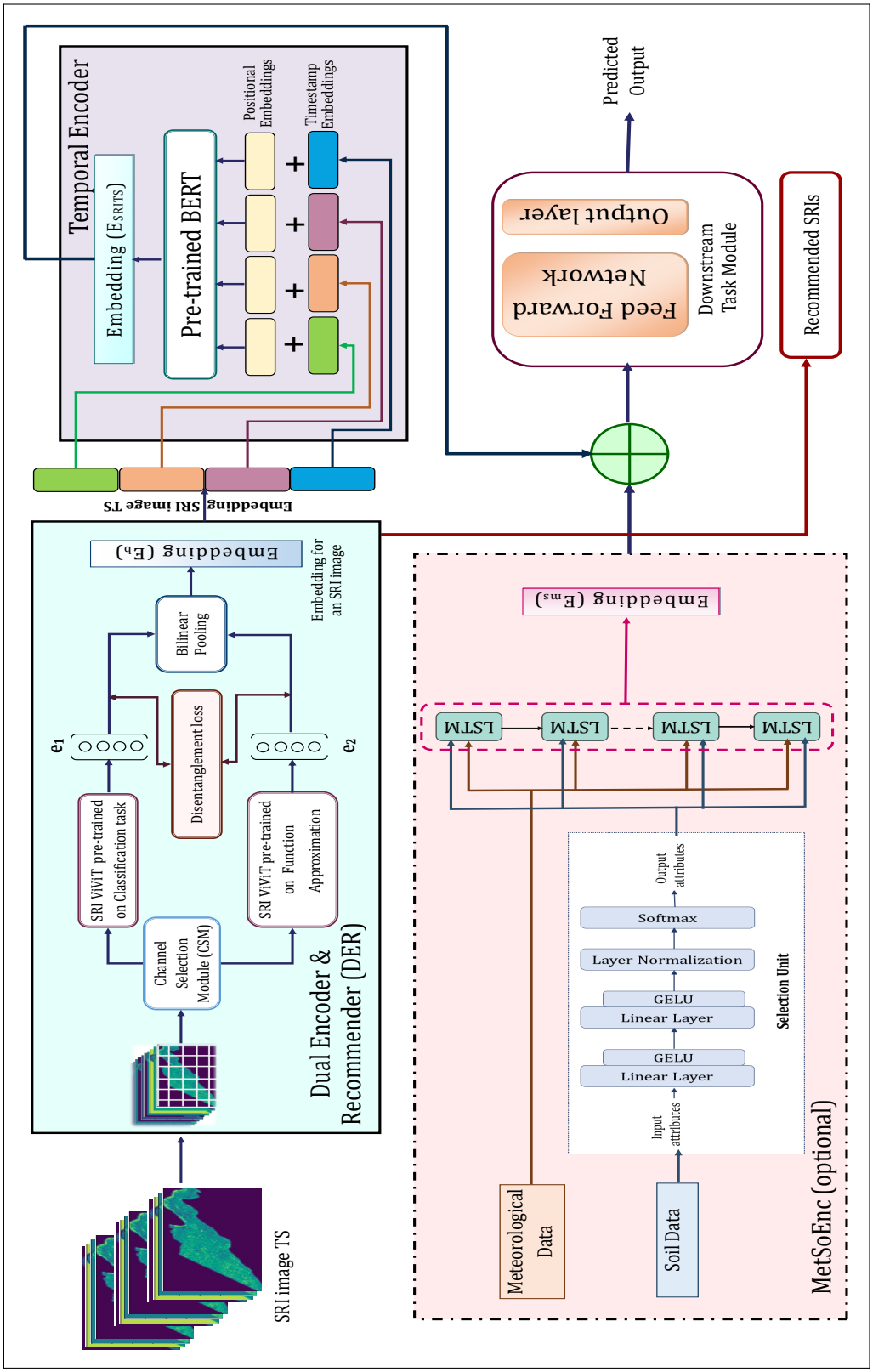

Figure 6.1: SpInN Architecture

6.4 Proposed Model: SpInN

We introduce a model, Spectral Indices Network (SpInN), which determines the relevance of different spectral reflectance indices for earth observation applications. The proposed model is depicted in Figure 6.1. It consists of four modules viz. - Dual Encoder and Recommender (DER); Temporal Encoder; MetSoEnc; and Downstream Task Module (DTM). SRI image at every timestamp is processed individually using DER which also recommends the spectral reflectance indices relevant to the application. The image embeddings at every time stamp are passed to the temporal encoder which learns the temporal patterns from the time series. The temporal encoder embeddings are forwarded to DTM for prediction or classification. MetSoEnc which is used to incorporate meteorological and/or soil data depending upon the application at hand. The module gives joint representations of the meteorological and soil data. This is an optional module and if used its embeddings are appended to the embeddings of the temporal encoder and forwarded to DTM.

6.4.1 Creating SRI Images

The raw images of MODIS have 5 bands - Red, Green, Blue, NIR, and SWIR. To create an SRI image with k channels, we calculate different spectral reflectance indices at every pixel of the satellite image and stack each SRI matrix them one after the other. (see Figure 6.2).

6.4.2 Dual Encoder and Recommender

Dual Encoder is designed to get a representation of spectral index images at every timestamp. It also effectively selects and recommends the most suitable spectral indices for the end task. It has three sub-modules described below:

6.4.2.1 Channel Selection Module (CSM):

CSM is designed to leverage relationships among different spectral reflectance indices within the SRI images. It utilizes average pooling and max pooling operations to effectively reduce spatial dimensions while retaining crucial channel (spectral reflectance

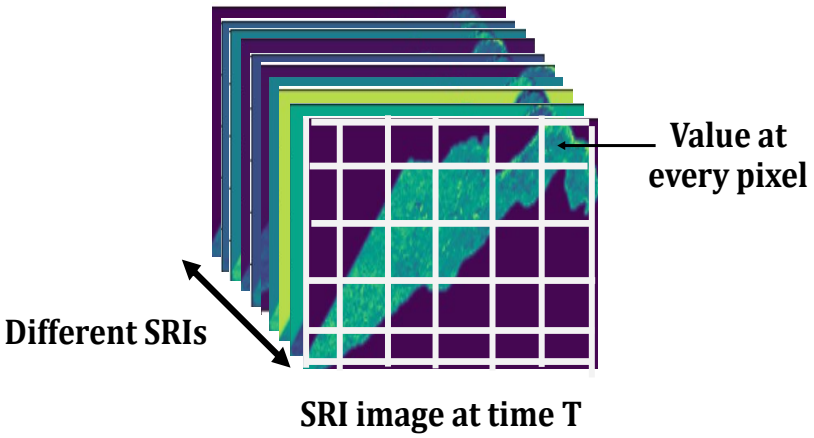

Figure 6.2: Spectral Reflectance Index Image

index) information. A two-layer feed-forward neural network is employed to model channel dependencies, incorporating ReLU activations. The output of FFN undergoes softmax yielding attention scores for each channel. The channel attention is an integral component of Dual Encoder, empowering the model to dynamically emphasize or de-emphasize the channels based on their contextual significance, thereby helping the model to capture task-relevant SRIs. The mathematical representation is:

$$X_{avg} = AvgPool(X) \tag{6.1}$$

$$X_{max} = MaxPool(X) \tag{6.2}$$

$$FC_{avg} = ReLU(Linear(X_{avg})) \tag{6.3}$$

$$FC_{max} = ReLU(Linear(X_{max})) \tag{6.4}$$

$$X_{combined} = FC_{avg} + FC_{max} \tag{6.5}$$

$$X_{attn} = softmax(X_{combined}, dim = 1) \tag{6.6}$$

$$X_{selchn} = X \times X_{attn} \tag{6.7}$$

where X is an SRI image, X_{attn} is the attention score of the channels, and X_{selchn} represents the modified SRI image after multiplying the attention score of each channel by its input matrix. X_{selchn} is passed to SRI ViViT for further processing.

6.4.2.2 SRI ViViT:

We adapted video vision transformer (ViViT) [34] designed for getting representations of RGB videos. In ViVIT, the image time series of dimension $Z \times H \times W$ is converted into multiple tubelets of dimension $z \times h \times w$ and tokens are extracted from the temporal, height, and width dimensions. Tubelet embeddings allow the model to efficiently capture spatiotemporal information by representing non-overlapping spatiotemporal patches. We use the factorized encoder version of the transformer which consists of two separate transformer encoders viz spatial encoder and temporal encoder. The spatial encoder, only models interactions between tokens extracted from the same temporal index. These representations are then forwarded to a temporal encoder.

We modified ViViT to make it work in conjunction with the channel attention module. The spatial features are exploited using the spatial encoder. If we use ViViT in its conventional way, it is not possible to retain the channel information separately. The model gives the collective embeddings of the entire video or SRI image. Our requirement is to get the features of individual channels to decide the importance of the channel for the considered task. Thus, we apply the temporal encoder of ViViT on the channel dimension for each SRI image individually to learn the features of every channel rather than applying it to the SRI time series. Thus, the transformer exploits the features of the channels which helps CSM to learn and identify the relevant channels.

We use two parallel SRI ViViTs, which we pre-train on two different tasks to make them exploit different features from the same SRI images (see section 5). We use disentanglement learning during fine-tuning to make sure that the two embeddings learn complementary features. The two embeddings (e_1 and e_2) obtained from SRI ViViT models are passed to the bilinear pooling layer.

6.4.2.3 Bilinear Pooling:

Bilinear pooling is mostly employed in computer vision, to combine the features from two different modalities or sources. Similarly, we use bilinear pooling to join the embeddings obtained from two SRI ViViT models. Bilinear pooling exploits the higher-order information and complex relationships in the two embeddings using pairwise correla-

tions. The final flattened tensor serves as a fused feature representation, encapsulating the joint information derived from both embeddings. Although bilinear pooling leads to redundancy to a certain extent [149], but we handled this problem using disentanglement learning. This forced the models to learn the complementary features and enhanced the model's ability to understand intricate cross-modal relationships and model performance across diverse objectives during fine-tuning. The layer gives the joint embedding e_{bt} at timestamp *t* given as follows:

$$e_{bt} = e_1^T A e_2 \tag{6.8}$$

where *A* is the weight matrix which plays an important role in shaping the interactions between the two embeddings. The final embeddings of all the timestamps, represented by e_b, are passed to the temporal encoder.

6.4.2.4 Temporal Encoder:

We use NLP transformer BERT [35] as a backbone in the temporal encoder. It uses the attention mechanism to exploit and learn long-term dependencies in the input sequence.

The SRI image embeddings are obtained at every timestamp and form SRI TS which are passed to BERT along with the positional embeddings. These embeddings are analogous to the embeddings of words in a sentence in NLP. We made two modifications to BERT to make it suitable for SRI TS. First, we replace the default embedding layer (which generates embeddings for text data) with dual encoder embeddings. Second, we have not used the "cls" token in order to get the task-agnostic embeddings. Without the *cls* token, the model learns more generic and intrinsic features from the data that apply to a broader range of tasks. We also pre-train BERT (see section 5) on the SRI time series. The mathematical representation of the module is as follows:

$$E_{token} = Embeddings(X),\ X\ is SRI\ image\ TS \tag{6.9}$$

$$E_{token} = E_b \tag{6.10}$$

$$E_{pos} = \begin{cases} sin(t/10000^{2i/d}) & i\%2 == 0 \\ cos(t/10000^{(2i+1)/d}) & i\%2 \neq 0 \end{cases} \tag{6.11}$$

$$Inp_{bert} = E_{token} + E_{pos} \tag{6.12}$$

$$E_{SRITS} = TSEnc(Inp_{bert}) \tag{6.13}$$

where *i* is the dimensions, E_{SRITS} represents the final embedding of the SRI time series.

6.4.3 Downstream Task Module (DTM)

DTM comprises a feed-forward network with two linear layers activated by the GELU activation function followed by the regression or the classification layer. GELU activation combines the properties of zone out, dropout, and ReLU for intensifying the probability of neuron output.

6.4.4 MetSoEnc

MetSoEnc is designed to incorporate and learn the combined representation of meteorological and soil data, wherever applicable. It contains an attribute selection unit [93] for soil attributes to select *j* most relevant attributes. The selected attributes are concatenated with the meteorological attributes and the combined set of attributes is passed to the bidirectional LSTM layer. If the application requires only meteorological data, soil attribute selection module is ignored.

$$E_{ms} = MetSoEnc([M_1, M_2, ...M_n], \tilde{A}) \tag{6.14}$$

where E_{ms} represents the embedding for meteorological (and soil attributes).

6.5 Learning Objectives

We use Euclidean distance to get a disentangled representation of two embeddings of the dual encoder, mean square error, and cross-entropy loss for prediction and classification, respectively.

Disentangled representation learning: Disentangled representation learning (DRL) is used to learn embeddings that capture different characteristics and are independent of each other. By using disentanglement learning, we ensure that each embedding captures a different aspect of the data. This property allows the representations to be used for various downstream tasks [150]. We used the concept for disentangling the embeddings obtained from SRI ViViT models pre-trained on different tasks. This leads to improved data quality and more reliable information extraction. Disentangling these embeddings ensures that these embeddings are generic, transferable, complementary, and capture different aspects of the data in a concise form which reduces redundancy. This facilitates improved performance of the model when adapting and fine-tuning to new tasks. The disentangled representations enable faster convergence and require less task-specific fine-tuning. It also mitigates task-specific biases that may arise during pre-training and reduces the risk of biased transfer when adapting the model to a new downstream task.

We calculate the Euclidean distance between the embeddings (e_1 and e_2) obtained from the two pre-trained SRI ViViT models as given in the equation below:

$$L_{drl} = ||e_{1_i} - e_{2_i}|| \tag{6.15}$$

Mean Square Error: We used mean square error as a loss function for the prediction downstream task. The mathematical representation of the loss function is given in the equation below where y_i and $\hat{y}_i$ are the actual and predicted output, respectively for a given location and year. *N* is the total number of location-year pairs.

$$L_{dt} = MSE = \sum_{i=1}^{N} \frac{(y - \hat{y}_i)^2}{N} \tag{6.16}$$

Cross entropy loss (CEL): This is a log loss function that we used in classification. In the equation below, y_i is the actual labels, $\hat{y}_i$ is the predicted labels, and *C* is the number of classes.

$$L_{dt} = CEL = \sum_{i=1}^{C} y_i.log(\hat{y}_i) \tag{6.17}$$

Total Loss: Amongst the two losses considered in training, downstream task (L_{dt}) needs to be minimized, and L_{drl} between the SRI ViViT embeddings needs to be maximized. The total loss in the model training is given in the equation below.

$$Total\ loss(L) = L_{dt} - L_{drl} \tag{6.18}$$

6.6 Pre-training

Taking inspiration from applications in computer vision and NLP we pre-train ViViT and BERT on spectral reflectance index images and SRI image time series, respectively. We pre-train ViViT for two pre-training objectives and BERT for one. The size of the dataset used for pertaining is 600GB. We used following pre-training taks:

- ***Function approximation task:*** This is a regression task in which a model is pre-trained to estimate the value of one attribute based on the values of other attributes in the dataset. We used it to pre-train ViViT using the set of 10 SRIs and estimate the value of another spectral reflectance index - Optimized Soil Adjusted Vegetation Index (OSAVI). The task learns the relationships among different SRIs and captures intricate patterns in the data. Exposure to noise in the function approximation task equips the model to handle uncertainties in satellite data, ensuring improved robustness.

- ***Is TS Ordered(Y/N):*** We also pre-train ViViT on the classification task to identify whether the given input is temporally in the correct order or not. The model is given as input the original SRI time series and the shuffled time series by randomly permuting the original time series. ViViT is applied iteratively on every image of the SRI time series. The task captures the contribution of each SRI image in the overall temporal context.

- ***Reconstruction task:*** We pre-train the BERT model using the next frame reconstruction pretext task. As the BERT model is designed to learn the temporal fea-

tures, we passed the embedding vectors of each time stamp of the spectral index time series obtained from ViViT models pre-trained on the function approximation task to the BERT model. We use an additional decoder to construct the spectral index image at the next timestamp from the predicted SRI vector.

Note: The ViViT models pre-trained on two different tasks are used in parallel in the disentangled representation learning setup to make sure that they are fine-tuned to extract different and complementary sets of features for the downstream task. Also, while pre-training BERT, the weights for the ViViT model are frozen. Using the pre-trained models in parallel helped in the concurrent utilization of their respective strengths. Each model may capture different aspects of the data, leading to a richer, more comprehensive, and holistic understanding. The redundancy provided by a parallel model contributes to robustness, especially in situations where one model may falter or exhibit biases. The parallel utilization of pre-trained models can facilitate transfer learning across related tasks, optimizing performance and generalization capabilities.

The loss function used for the function approximation task and the reconstruction task is the mean squared error and Binary Cross entropy loss for the classification task.

6.7 Models for Comparison

Out of six applications considered, existing studies have used spectral reflectance indices only for crop yield prediction. We conducted a comprehensive comparison of our proposed model SpInN with four existing models working with SRIs viz., Linear Regression (LR) [32], Random Forest regression (RFR) [33], Support Vector Regression (SVR) [32], and LSTM [32]. These models predict crop yield using SRIs as a single numeric value for a specific location (county). We also compare our model with four other existing models viz. CNN[43], CNN +GP [43], CNN+LSTM [44], and CYN [84], working with histogram time series. To ensure fair comparisons, all the existing models were optimized and trained for the same location and year as used in training SpInN. For other prediction applications, we chose LSTM (best-performing model using SRIs) as the baseline

for comparison (see Table 6.3). For landcover classification, we compared SpInN with Resnet50 [18].

Thus, the models selected for this comparative analysis are divided into three types:

6.7.1 Models for prediction problems working on SRIs

The details of models used for prediction problems are as follows:

- **Linear Regression:** Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. Most of the researchers have used linear regression using only one spectral reflectance index to predict crop yield. A few of them have used multiple linear regression for crop yield prediction using 2-3 spectral indices. We have applied multi-linear regression using 10 SRIs for crop yield prediction.

- **Random Forest Regression:** Random Forest Regression is an ensemble method that uses different independent decision trees to make predictions. The final prediction is obtained by aggregating the predictions of all the trees using the average method. Random Forest Regression captures non-linear relationships and provides stable and accurate predictions. After hyper-parameter tuning, we used the number of base estimators as 37 and the maximum depth for each decision tree as 11.

- **Support Vector Regression:** SVR operates by identifying a hyperplane that best represents the distribution of data points in a high-dimensional space, where the distance between the hyperplane and the data points is minimized. We have used the Radial Basis Function (RBF) kernel which helps in capturing the non-linear relationships between input features and target variables. It helps SVR to capture intricate patterns and dependencies in the data that linear models are not able to. After hyper-parameter tuning, we have used epsilon as 0.03.

- **Long Short Term Memory:** LSTM is a type of recurrent neural network (RNN) with forget and memory gates designed for capturing long-term dependencies in

sequential data. These gates help LSTM to decide when and what information needs to be forgotten. The gates and the cell states in LSTM make it suitable for dealing with the long-term dependencies and problem of vanishing gradient in RNN. We trained and optimized LSTM using Adam optimizer keeping a learning rate of 0.0001 after hyper-parameter tuning.

6.7.2 Models for prediction problems working on histograms

The next set of models are working with histogram time series. To work with histograms, researchers have created a separate histogram for each reflectance band by aggregating the pixel intensities. Thus histograms represent the count of pixels in a certain intensity range. The details of the models are:

- The researchers [43] applied CNN and CNN+GP models on histogram time series of reflectance bands obtained from MODIS for crop yield prediction. CYP is modeled as a static problem and they did not exploit the temporal patterns in the data, which is crucial for appropriate modeling of crop yield prediction problem.

- CNN+LSTM [44] uses histograms of MODIS reflectance bands and soil data. The authors apply CNN over soil histograms to model soil as static data, and LSTM for learning temporal patterns in the surface reflectance data.

- CYN [84] models crop yield prediction as a spatiotemporal problem, also incorporating the depth-sensitive information of soil data. CYN models soil data to select the required depth level for each soil attribute at every timestamp of the crop cycle.

6.7.3 Model for Land cover classification:

The above-mentioned models are only suitable for prediction problems. To validate the performance of the proposed model SpInN in a classification problem, we compared it with an existing model ResNet used in study [18]. The authors have not given the details for which version of the model they used, so we opted for ResNet50 with four blocks.

6.8 Experimental Setup and Evaluation Metric

This section covers the details of the experimental setup and valuation metrics used.

6.8.1 Experimental Setup

All the experiments are performed using Pytorch 1.11.0 and CUDA 11.7 on an A100 GPU server having 80GB GPU RAM. All the models are trained using the Stochastic Gradient Descent optimizer for 50 epochs with a batch size of 16. The momentum and weight decay are 0.3 and 0.001, respectively for all the applications. The learning rate η for the optimizer is decided while tuning hyper-parameters for each model and application. For CYP, CCP, SMP, SEP, SCP, $\eta = [0.0003, 0.001, 0.001, 0.0001, 0.00001]$, respectively. We tested the models for two years (2019 and 2020). All the predictions are done at timestamp $t + 1$, taking 1 to *t* as input TS data. If prediction needs to be done yearly, we averaged the predicted output for the years 2019 and 2020. If the application requires prediction at a monthly level (for SMP and SCP), the average is taken for 24 predictions in two years, and so on. We have performed the experiments five times and given the best-observed results in the results section.

6.8.2 Evaluation Metrics

We used two key metrics viz. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for prediction problems. For classification, we used accuracy and f1-score.

Evaluation metrics for prediction: We used two key metrics for prediction problems viz., root mean square error (RMSE) and mean absolute error (MAE) for evaluating and comparing the performance of our proposed model. RMSE calculates the root mean squared magnitude of errors, assigning more weight to larger errors and MAE provides insight into the mean magnitude of errors in the model predictions. A higher RMSE implies that the model is encountering a notable quantity of errors with substantial magnitudes. It indicates that the model is struggling with a considerable number of errors characterized by significant magnitudes. Its tendency to heavily penalize large errors due

to the squaring operation makes it sensitive to outliers. The formula for both the metrics is given below:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_{c,T}^{i} - \hat{y}_{c,T}^{i})^2}{N}} \tag{6.19}$$

$$MAE = \frac{\sum_{i=1}^{N}|y_{c,T}^{i} - \hat{y}_{c,T}^{i}|}{N} \tag{6.20}$$

where N is the total number of $\{c, T\}$ pairs, c is the county, T is a year for crop yield prediction, a month, for snow cover and soil moisture prediction, and for solar energy and cloud cover prediction it is a fortnight.

Evaluation metric for classification: We used accuracy and f1-score as the evaluation metrics. Accuracy measures the overall correctness of a classification model. F1-score combines precision and recall into a single value. It provides a balance between precision and recall, making it useful in imbalanced data. The formula for both the metrics is given below:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Predictions} \tag{6.21}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6.22}$$

6.9 Results and Discussion

Comparison with existing models on crop yield prediction: The first set of experiments is performed for crop yield prediction. Table 6.2 shows RMSE and MAE obtained by different models using all 10 listed spectral reflectance indices. It can be observed from the table that the proposed model outperforms the existing models. There is a significant reduction in RMSE obtained by SpInN. The RMSE is reduced for corn yield prediction by

Table 6.2: RMSE and MAE for CYP

Model	Corn		Soybean	
	RMSE	MAE	RMSE	MAE
LR [32]	37.3733	27.1181	8.3813	7.5609
RFR [33]	23.7021	21.9117	7.3676	6.9295
SVR [32]	27.8096	25.5323	9.9320	9.6108
LSTM [32]	22.1133	18.8315	6.0310	5.7554
CNN [43]	27.1690	25.1255	12.436	10.0146
CNN+GP [43]	28.2040	26.6421	10.115	8.5051
CNN+LSTM [44]	23.5860	21.3289	9.9560	7.6278
CYN [84]	21.5050	18.1359	8.1660	5.8712
SpInN	**18.8498**	**17.1664**	**5.9669**	**5.5609**

approximately 37%, 19%, 32%, and 14% in comparison to LR, RFR, SVR, and LSTM. Similar results are obtained for soybean yield prediction where a minimum reduction in error is observed in comparison to LSTM 5.49% and the maximum is observed from SVR which is 42.6%. The percentage improvement of approximately 12% and 27% is observed, respectively for corn and soybean yield prediction when compared with the best-performing histogram model, CYN. The superior results obtained by SpInN can be attributed to multiple factors including the use of SRI images in place of single aggregated value (as used in existing models of SRIs) and the use of transformers in the architecture.

Channel Selection: The next set of experiments is performed to find the importance of different spectral reflectance indices for the end tasks. The attention for each channel/spectral index is calculated using the CSM of the dual encoder. The graphs depicting the weight of each SRI for all the applications are given in Figure 6.3. We have used the inverse-scree method [71] and the weight value of SRI to select the spectral indices relevant to the given application. For example for corn, soybean, soil moisture, and cloud cover we recommend 5 spectral indices. For solar energy and classification, we recommend 6 SRIs and for snow cover prediction, 7 channels are recommended. The selected SRIs are taken from left to right before the red mark in each graph.

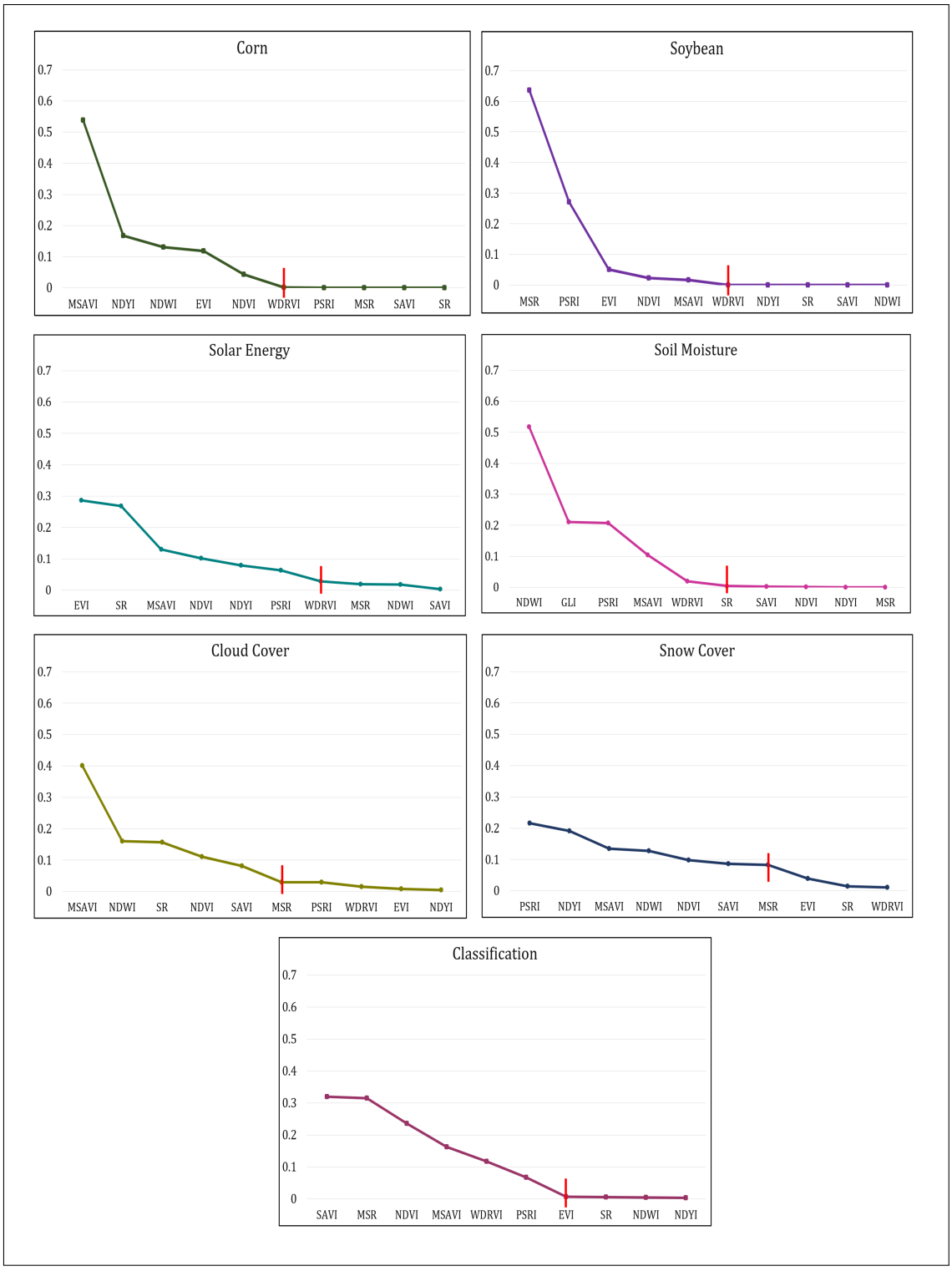

Figure 6.3: SRI Selection

Table 6.3: Impact of using SRIs relevant to applications (RMSE)

Application	10 SRIs		Selected SRIs	
	SpInN	LSTM	SpInN	LSTM
CYP(Corn)	18.849	22.113	17.446	20.075
CYP(Soybean)	5.6996	6.0310	5.1647	5.8607
SCP	4.7837	9.6103	4.1836	8.3642
SEP	6.2235	8.6349	5.7313	7.8735
SMP	5.1203	10.677	4.3802	9.6408
CCP	13.898	23.347	13.519	21.758

Significance of SRI recommendation: The next set of experiments is performed to see the impact of using only recommended SRIs for each application. We performed the experiments to validate the significance of recommended SRIs for SpInN and the best performer amongst the existing models i.e. LSTM for all applications. It can be observed from Table 6.3 that using only the recommended spectral reflectance indices significantly affects models' performance. The RMSE obtained using SpInN for corn yield prediction is 18.849 bu/ac when using all the spectral indices and it is reduced to 17.449 bu/ac when using only the set of recommended five spectral indices ($\approx$ 7.4% improvement). Similarly, using the LSTM model the RMSE reduced from 22.113 bu/ac to 20.075 bu/ac ($\approx$ 9.2%). In the case of snow cover prediction, RMSE reduced by $\approx 12\%$ on using selected spectral indices for both models. Similar results are obtained for other applications. The reduced error using selected SRIs signifies that irrelevant SRIs possibly add noise to the data and thus lead to degraded performance. Some other benefits of using the reduced number of spectral indices include reduced computational and memory requirements during training and inference.

Significance of Additional data: Table 6.4 contains the output of experiments conducted using additional meteorological and/or soil data depending upon the requirement of the application. We conducted experiments for the existing LSTM model and the proposed model SpInN. It is clear from the table that incorporating meteorological and soil data has a significant impact on all the applications. Incorporating meteorological

Table 6.4: Impact of additional data (RMSE)

Application	Only selected SRIs		SRI+additional data	
	SpInN	LSTM	SpInN	LSTM
CYP(Corn)	17.446	20.075	14.4796	18.0675
CYP(Soybean)	5.1647	5.8607	4.5488	5.3894
SCP (no soil)	4.1836	8.3642	4.0248	8.0104
SEP	5.7313	7.8735	5.505	7.2837
SMP	4.3802	9.6408	3.6209	9.1440
CCP (no soil)	13.519	21.7585	10.8465	19.242

Table 6.5: Ablation Study: Significance of DRL and BP (RMSE)

Application	SpInN	SpInN-DRL	SpInN-DRL-BP
CYP(Corn)	14.479	14.701	15.197
CYP(Soy)	4.5483	4.8783	5.0901
SCP	4.0248	5.0727	5.3232
SEP	5.5050	5.6326	6.2783
SMP	3.6209	4.3208	4.4964
CCP	10.846	11.003	11.257

and soil data has improved the corn yield predictions by approximately 17% and 10% using SpInN and LSTM models, respectively. Similar patterns were obtained for other applications. For example, RMSE is reduced by $\approx 4\%$ for snow cover prediction by incorporating meteorological data in both models.

Significance of Disentangled representational learning and Bilinear Pooling: We perform experiments without using the disentangled representation learning and bilinear pooling to verify their significance. SpInN represents the model when DRL is removed from the model. It is observed that the model is not able to handle redundancy introduced due to bilinear pooling, and the model is also not able to learn complementary features. Both these problems adversely affect the RMSE as shown in column 3 of Table 6.5.

To emphasize the importance of using two parallel pre-trained SRI ViViT models in parallel, we removed both bilinear pooling and disentanglement learning from SpInN.

Table 6.6: Comparison: Landcover classifcation

Model	10 SRIs		Selected SRIs	
	SpInN	ResNet	SpInN	ResNet
Accuracy	84.56%	42.82%	89.9%	47.29%
F1-score	0.595	0.539	0.612	0.557

SpInN-DRL-BP represents this variant of the model. In this scenario, we can use only a single pre-trained SRI ViViT model. This results in a significant increase in RMSE for all the applications. It can be observed from Table 6.5 that there is a significant impact on RMSE for all the applications considered (see column4, Table 6.5). The most adversely affected application is snow cover (increase in RMSE by approximately 24%), followed by soil moisture with an increase of 19% in RMSE.

Results for Land Cover Classification: We achieve an accuracy and f1-score of 84.56% and 0.595, respectively when using all the SRIs for classifying the land cover into 17 classes. The accuracy and f1-score increased to 89.9% and 0.61, respectively when we used only six relevant SRIs recommended by SpInN. It can be observed from Table 6.6 that SpInN outperforms Resnet.

Except for CYP, there are only a few studies [18, 20, 98, 103, 142, 144, 146] available for other applications considered. Most of the existing solutions related to these applications are confined to very small regions and they address only one application. Except for the proposed SpInN, there is no existing model which generalizes for multiple applications. The existing models are either shallow machine learning models or simple neural networks which are not capable of learning highly non-linear patterns in satellite data.

Standard Deviation in results: All the results mentioned are best-case results. However, we performed each set of experiments five times and observed a standard deviation of a maximum of 0.3 for all applications. The details are given in Table 6.7.

6.10 Main Contributions

The main contributions of the chapter are:

Table 6.7: Standard deviation in experiments for SpInN

Application	10 SRIs			Selected SRIs		
	Best-case	Mean	Std-dev	Best-case	Mean	Std-dev
Corn	18.8498	19.1185	±0.1830	17.4495	17.7172	±0.2721
Soybean	5.6996	5.8844	±0.2017	5.1647	5.2473	±0.0821
SCP	4.7837	4.9791	±0.1682	4.1836	4.3232	±0.0972
SEP	6.2235	6.6378	±0.3070	5.7313	5.9100	±0.1704
SMP	5.1203	5.4330	±0.2362	4.3802	4.5792	±0.2169
CCP	13.8985	14.4628	±0.3302	13.5196	13.7501	±0.2089

1. We explore the use of spectral reflectance indices in different earth observation applications and introduce a model to recommend the most relevant SRIs for an application. We have applied and validated the recommendations for six different applications.

2. We innovatively introduce a channel attention mechanism applied to features obtained from the temporal encoder of SRI ViViT which is employed on channels of SRI image. The time series of SRI images is then exploited by BERT for temporal patterns.

3. We pre-train ViViT on SRI images to learn their spatial-spectral features and BERT on a time series of features obtained from ViViT (representing SRI time series) to learn temporal features in the SRI time series.

4. This is the first attempt to model spectral reflectance indices as an image to preserve the SRI properties in the spatial modality.

6.11 Summary

We attempted to establish that spectral reflectance indices and their sensitivity towards the detection of vegetation, environmental variables, and physiological and morphological characteristics are important and can be used for various earth observation applications. We also show that a subset of SRIs is sufficient to solve an earth observation application.

We have innovatively modeled prediction and classification problems using a novel concept of SRI images, unlike previous existing studies which use single value representation of SRIs for a large region like a county for a time stamp. This leads to the loss of critical spatial information. We have used the image representation to address the spatial information loss problem. We also propose a model, SpInN which can recommend relevant SRIs for a given application and solves the problem using the subset of recommended SRIs.

SpInN uses ViViT and BERT innovatively on SRI image time series to learn its spatial-spectral-temporal representation. Our results demonstrate that SpInN gives state-of-the-art results for six earth observation applications considered. The proposed architecture is generic and can be used for many other earth observation applications. The pre-trained dual encoder of SpInN can be used independently for applications that require only static SRI image analytics. SpInN automates the spectral reflectance index recommendation process for various applications, which otherwise require the services of domain experts.

Chapter 7

SaTran: A transformer for Satellite Image Time Series

[1]

7.1 Introduction

Satellite Image Time Series (SITS) data offer valuable insights into Earth's surface characteristics and dynamics. It has widespread applications across domains like ecology, agriculture, forestry, land management, disaster monitoring, risk assessment, etc. Deep learning has gained popularity in the remote sensing community due to its ability to learn valuable features from input data without feature engineering. In the realm of SITS classification, a combination of convolutional and recurrent neural networks is employed to capture spatiotemporal characteristics from the data. An alternative to RNNs, transformers, originally proposed for natural language processing tasks, have shown promising performance in sequence encoding. A couple of studies [25, 26] used BERT to classify

[1] *The work presented in this chapter is communicated:*

- *Arshveer Kaur, Poonam Goyal, Niranjan, and Navneet Goyal, "SatTran: A transformer for high spatial resolution satellite image time series exploiting spatiotemporal redundancies", in NeurIPS 2024. [Communicated]*
- *Patent filing in process.*

time series for every pixel which limits them from effectively exploiting the spatial correlations in the image time series. Moreover, these models work only for classification as they segment the image time series which is not suitable for prediction problems like prediction of crop yield, snow cover, cloud cover, etc. In prediction problems, the ground truth is mostly available for coarser granularity than that of pixel e.g., at a county or a district level. Another model TSViT [24] used ViT for landcover classification. The authors factorize input dimensions into spatial and temporal components to reduce the computation. However, the model is not able to identify the redundancy in the patches and processes them all.

Satellite data is much larger than the large datasets used in NLP or vision domains [151]. The size of an image for Landsat-8 satellite, a high spatial resolution satellite, is approximately $2000 \times 2500 \times \#bands$. For yearly time series of Landsat-8, the size becomes $2000 \times 2500 \times 5(\#bands) \times 23(\#timestamps)$ which makes its data volume $\approx$ 700MB. The total data used in the study for pre-training and downstream tasks comprises of around 2000 counties making the amount of data is 10.0 TB for Landsat-8 which is much larger than the datasets used in other domains [152]. Processing single image time series of Landsat-8 using existing video vision models leads to GPU out-of-memory (OOM) error on A100 GPU card with 80GB memory.

There are a few models like AdaVit [153], and DynamicVit [154] which reduce the computational requirements by processing images in patches and using attention to select only informative patches. Authors [155] used the token dropping mechanism to ignore less informative tokens while pre-training BERT. However, as mentioned Esther Rolf et. al. [151] any models developed for 3-channel RGB images or videos are suboptimal for modeling satellite datasets due to the large volume and different characteristics of SITS data in comparison to RGB data. Thus, we require a method which can train a Large SITS Model efficiently.

We propose a transformer model, SaTran, for large size satellite image time series which exploits spatiotemporal redundancies. SITS data can be characterized by the presence of patches with spatiotemporal redundancy persisting throughout the time series,

referred to hereafter as redundant patch tubes. SITS data also contains patches where temporal redundancy lasts only for a few timestamps, referred to hereafter as non-redundant patch tubes. The pictorial representation of the classification of patch tubes is given in Figure 7.1. For example, a region of a barren land/water body has spatiotemporal redundancy, and it won't change even for years (thus is a redundant patch tube); 2) the non-redundant patches (regions of interest) experience changes with time but can still have a temporal redundancy for a shorter span, for example, cultivation land experiences changes in the crop cycle duration. However, during harvest time when the crop is fully grown, there can be redundancy for a few time stamps. Removing redundancies reduces the computational requirements thereby helping in the democratization of satellite image technology.

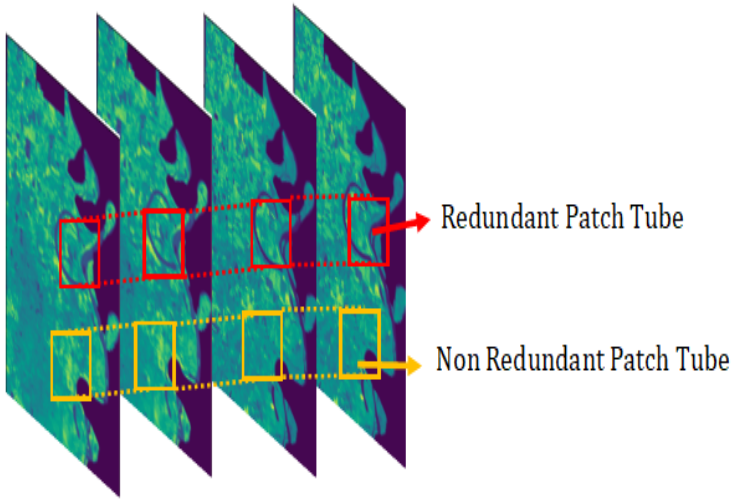

Figure 7.1: Classification of Patch Tubes

SaTran disentangles spatiotemporal and temporal redundancies and makes the SITS processing efficient. It has two modules - PatchTubeSelect and TemporalRedundancyHandler. We first remove spatiotemporal redundancies with the help of PatchTubeSelect which selects hotspots (non-redundant patch tubes) using an attention mechanism to discern critical areas necessitating focused attention and exclude the redundant patch tubes. We, then use TemporalRedundancyHandler which innovatively uses VideoMAE [156] on

non-redundant patch tubes to further handle temporal redundancy local to these patches. The twofold redundancy handling approach of SaTran helps mitigate the impact of noise and data leakage leading to a more accurate representation of the SITS.

7.2 Related Work

Satellite Data: Advancements in satellite image technology are leading to notable enhancements in capturing data with high spatial, temporal, and spectral resolutions. This progress enables effective monitoring of the Earth's surface at desired levels of detail, catering to various Earth observation applications. Popular satellite systems are MODIS [10], Landsat-8/9 [40], and Sentinel-2 [41] due to their publicly available data which can be used in different real-world applications like disaster management, urban planning, agriculture, climate studies, etc. This makes satellite imagery a cost-effective solution for obtaining global-scale data. Time series analytics are commonly required for many applications leveraging satellite imagery. Satellite image time series is analogous to RGB videos but it has characteristics that differentiate it from RGB videos e.g. SITS do not contain moving objects, the landscape does not change their position and only the changes are observed in them with time. This is the reason why the existing video models are not suitable for SITS data.

Earth Observation Applications: Researchers have tried to model SITS data in different forms according to specific applications. They have used satellite data in the form of histograms [43, 44, 58], spectral reflectance indices [32, 33] or images [18] for different earth observation applications. Different studies have applied various statistical [129] and machine learning [46, 130–133] on different types of data for crop yield prediction.

Transformers in other domains: Transformers are gaining success in areas of NLP and vision. BERT [35] is a revolutionary text model that has significantly advanced the field of language understanding. Similarly, Video Vision Transformer (ViViT) [34] represents a significant advancement in the domain of computer vision, extending the transformer architecture to handle sequential frames of video data. VideoMAE [156] customized video tube masking approach characterized by an exceptionally high masking

ratio to extract effective video representations during the pre-training process. Authors [157] introduced making the decoder of VideoMAE and proposed model VideoMAE:v2 to improve the accuracy of video analytics. Lingchen et al. [153] proposed an Adaptive Vision Transformer (AdaViT) which processes an image in patches. It learns which patches to use and which self-attention heads to activate for every image and thus reduces the computation cost. Similarly, Rao et al. proposed DynamicViT [154] which prunes redundant tokens dynamically based on the input. The model divides the RGB images into independent patches and uses the attention masking concept to mask the tokens of patches which are of minimum importance. In another study [155], Hou et al. proposed the token-dropping concept for accelerating the pertaining process of BERT. These models can be explored for their use in SITS analytics but are not directly applicable to satellite data.

Models in SITS Analyisis: As stated in [151], the basic properties of two types of data- regular RGB videos and SITS are different. Unlike RGB videos, the objects in temporal sequences of satellite images remain in a fixed position but change in appearance over time. Researchers [25, 26] tried to adapt BERT for SITS classification at pixel level. Authors [25] presented a self-supervised pre-training approach of BERT designed to initialize a transformer-based network. The model is tasked with predicting randomly contaminated observations within an entire time series of a pixel. Similarly, authors [26] extended the above work to apply BERT to the time series of the immediate neighboring pixels and then predict the label of the center pixel using SITSFormer. The use of BERT on pixel time series shown to be a potential method for improving SITS classification performance and mitigating overfitting challenges in the application. Tarasiou et al. [24] proposed Temporo-Spatial Vision Transformer (TSViT) which is based on famous ViT model and adopted for satellite image time series analytics. TSViT divides a SITS into non-overlapping patches across both spatial and temporal dimensions which are then tokenized and processed by a factorized temporal-spatial encoder. However, these models are not suitable for applications like prediction of crop yield, snow cover, cloud cover, etc. where the ground truths cannot be available at pixel level. For such applications, the ground truths are available at a bigger region like a county or district and it necessitates

processing the data at the image time series level and not at pixel time series level.

The existing models for SITS are designed specifically for classification problems and are not able to solve prediction tasks. The video analytics models are not suitable to be adapted for SITS. The proposed model in this paper works to resolve these two problems.

7.3 Study Area and Data Used

The data used for SaTran consists of satellite image time series from two satellites MODIS and Landsat-8. The ground truth data is specific to the application. We have taken around 2000 counties from different states of US for all applications considered. The details are as follows:

- **Data Used:** For MODIS, time series for one year is $200 \times 250 \times 5(\#bands) \times 46(\#timestamps)$ and for Landsat-8, it is $2000 \times 2500 \times 5(\#bands) \times 23$ $(\#timestamps)$. For a fair comparison, we have used data for the same years (2014-2020) for both satellites and used only the bands common to both satellites i.e. Red, Green, Blue, NIR, SWIR.

- Crop Yield Prediction (CYP): The crop yield data for U.S. counties utilized in this study is taken from Quick Stat, a comprehensive database compiled by the United States Department of Agriculture (USDA) [80]. The data used in this study spans the period from 2002 to 2020. The yield values are quantified in bushels per acre (bu/ac), offering a standardized unit for assessing and comparing crop productivity across different regions. We used top producer counties of soybean from states - Michigan, North Dakota, Arkansas, Indiana, Tennessee, Ohio, South Dakota, Iowa, Kansas, Kentucky, Minnesota, Mississippi, Missouri, Nebraska, Illinois, and Wisconsin.

- **Snow Cover Prediction (SCP):** The snow cover data has been acquired at the county level using the MODIS product MOD10A1. This dataset offers information about snow cover extent with a spatial resolution of 500 meters using normalized

difference snow index (NDSI). NDSI is a key indicator, representing the percentage of the area covered by snow. We consider 300 counties from states experiencing annual snowfall of more than 250 inches. The counties lie in the states of New Hemisphere, Washington, Oregon, California, Colorado, and Utah.

- **Soil Moisture Prediction (SMP):** The soil moisture data is acquired at the county level for every month using NASA-USDA Enhanced SMAP Global soil moisture data [147]. This dataset was developed by the Hydrological Science Laboratory at NASA's Goddard Space Flight Center in cooperation with USDA Foreign Agricultural Services and USDA Hydrology and Remote Sensing Lab. The Soil Moisture Active Passive (SMAP) instrument measures the amount of water in the surface soil everywhere on Earth. The value represents the amount of water in mm. The data is available from 2016 and thus data used in this study is for 5 years from 2016 to 2020. We consider 275 counties from the states- Iowa, Kansas, Illinois, Kentucky, and Indiana.

- **Solar Energy Prediction (SEP):** The information about solar energy produced in a county has been acquired from visual crossing [105]. The value represents the total solar energy produced in MJ/m2 for a county in a day. The data used in this work spans the period from 2014 to 2020. The counties considered lie in the states- Iowa, Kansas, Illinois, Kentucky, and Indiana.

- **Cloud Cover Prediction (CCP):** Cloud cover represents the proportion of the sky covered by clouds throughout the day. We acquired data captured on a daily basis from visual crossing [105]. The data used in this study spans the period from 2014 to 2020. The counties considered lie in the states- Iowa, Kansas, Illinois, Kentucky, and Indiana.

- **Land Cover Classification (LCC):** Land cover can be classified into a different number of classes. We used MCD12Q1 [148] which classifies land cover into 17 classes. The dataset provides annotations for the land cover at a spatial resolution

Table 7.1: Length of time series used for different downstream applications

Satellite	CYP	SMP	SCP	SEP	CCP	LCC
MODIS	32	12	12	24	24	46
Landsat-8	16	6	6	12	12	23

of 500m and yearly granularity. The data used in this study spans the period from 2014 to 2020. The counties considered belong to states- Iowa, Kansas, Illinois, Kentucky, and Indiana.

7.3.1 Deciding length of time series:

The length of time series for an instance for every application is different due to the difference in the visiting frequency of satellite systems. For e.g. in snow cover and solar energy prediction the time series considered is from the last three months and the prediction is for the next month for snow cover and the next fortnight for solar energy. The details are given in Table 7.1. Other details of data pre-processing are given in Appendix A.

7.4 Proposed model: SaTran

In this section, we discuss the characteristics of satellite image time series and how it is different from RGB videos, and then describe the architecture of SaTran to handle challenges posed by SITS.

7.4.1 Characteristics of SITS

Satellite image data are stored in a raster format, organized as a tensor with dimensions for height, width, and channels. Temporal aspects can be integrated by arranging spatially-aligned rasters along a fourth dimension. Although this structure resembles images\videos, satellite images are far different from their equivalent in natural images for many reasons including:

1) Number of channels in satellite images and the size of the images from high spatial resolution satellites.

2) Unlike RGB videos, SITS often exhibit spatiotemporal redundancy over time, especially in specific landforms. For example, water bodies will have consistent patterns year after year. The spatial arrangements of landforms in SITS data do not change drastically over time. In contrast, RGB videos capture dynamic scenes where spatial configurations change frequently. The stable patterns observed in SITS contrasts with the fluid nature of video imagery, especially in urban environments or areas with constant human activity leading to huge spatiotemporal redundancy.

3) The rate of temporal redundancy is different in different spatial landforms. The redundant patch tubes such as water bodies, urban area, etc. will have temporal redundancy for a longer time. However, non-redundant patch tubes like areas under snow experience different redundancy for a shorter span. In regions where snowfall is a regular occurrence, such as high-altitude areas or northern latitudes, the temporal redundancy in satellite image time series can be less during the initial phase which can eventually be relatively high during winter months once snow settles and decreases as temperature rises. This leads to dynamic changes in the landscape. The once uniform snow cover gives way to patches of melting snow, revealing underlying terrain and vegetation. During this transition period, the temporal redundancy in satellite imagery decreases as the spatial patterns evolve rapidly.

4) SITS do not contain moving objects and thus the orientation of regular RGB videos is of more importance and SITS don't have any natural orientation [151].

7.4.2 Model Architecture

SaTran consists of two important modules – PatchTubeSelect and TemporalRedundancyHandler. The architecture is given in Figure 7.2. PatchTubeSelect handles the spatiotemporal redundancy in SITS by selecting and processing the patch tubes using an attention mechanism. Temporal Redundancy Handler uses VideoMAE to handle the local temporal redundancy of the selected patch tubes. Both modules are described below:

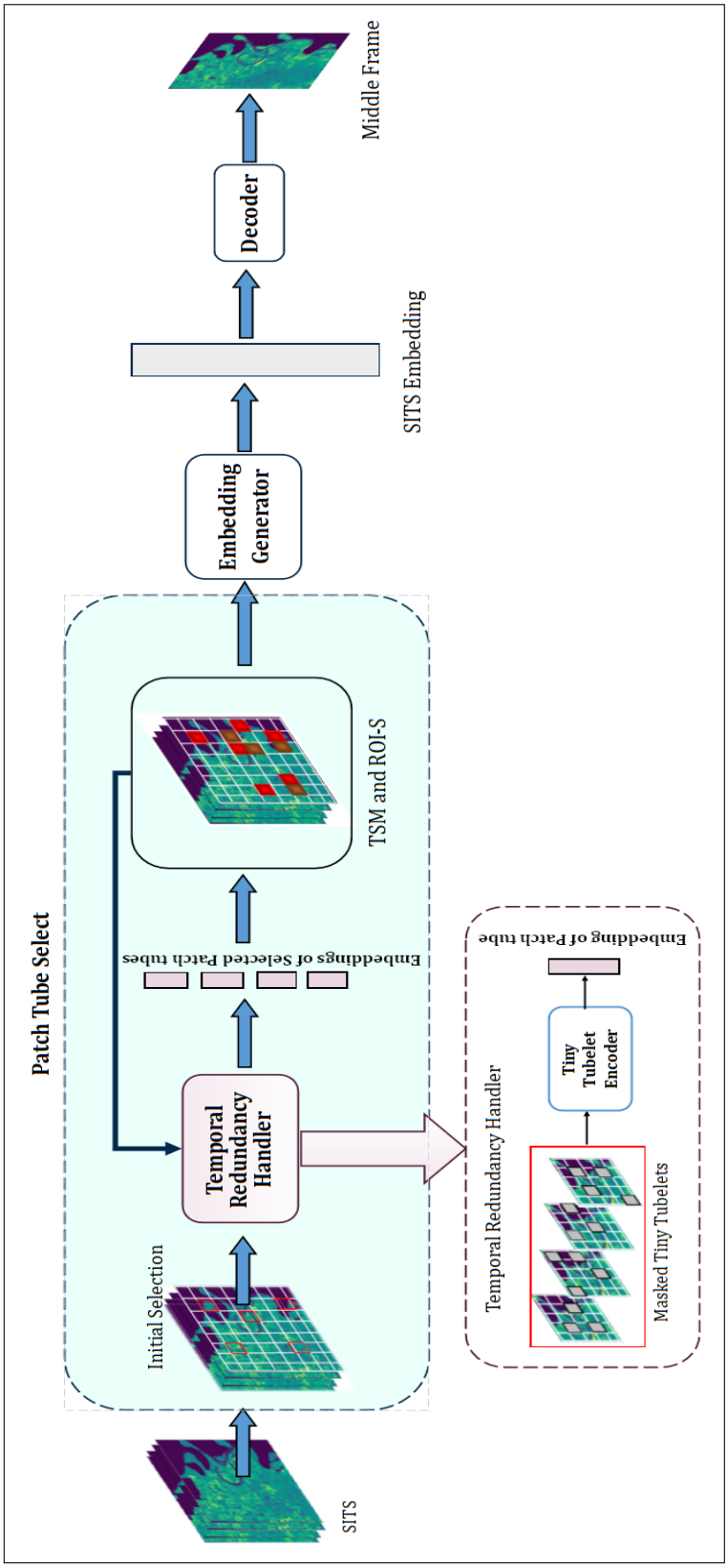

Figure 7.2: Model Architecture: SatTran

7.4.2.1 PatchTubeSelect:

A satellite image is gridded virtually into spatial patches. each patch consist of $n \times n$ pixels. A patch tube extends a patch in the temporal dimension. The patch tubes are passed to the Temporal Redundancy Handler which gives the representations of Patch tubes which are then forwarded to the Tube Selection Module (TSM) which is a sub-module in patch tube select. TSM utilizes attention scores to identify the top '*k*' tubes, which are then passed to the Region of Interest Selector (ROI-S) sub-module. ROI-S determines the unprocessed neighboring tubes of the top '*k*' tubes and generates a list of tubes to be processed in the subsequent iteration. Most of the neighboring patch tubes are excluded due to spatiotemporal redundancy and new patch tubes are randomly selected which then become the new regions of interest. This process iterates until a fraction (1/*x*) of the SITS is processed. The enhanced tube representations obtained from TSM are subsequently forwarded to the embedding generation module, which outputs the embedding of the entire SITS learned by the network across multiple iterations.

7.4.2.2 TemporalRedundancyHandler:

It adapts VideoMAE for handling temporal redundancy. The patch tubes are divided into tiny tubelets of size $(T/t \times H/h \times W/w)$ where *t, h, w* depends on the satellite used and details are given in section 6.1. These tiny tubelets are randomly masked and the remaining tubelets are processed using vanilla Vision Transformer (ViT) [158]. VideoMAE takes these tiny tubelets as inputs and uses the joint space-time cube embedding to obtain tubelet tokens. This reduces the volume of input data due to reduced spatial and temporal dimensions thus making the processing efficient. As the temporal redundancy in SITS persists for shorter lengths, the original high masking (95%) ratio in VideoMAE cannot be used for SITS. Also, there is less risk of information leakage in SITS, due to which we recommend reducing the masking ratio to 75% which is still significant and thus only a small number of tokens need to be processed by the encoder which further makes the process efficient. The comparative study for different masking ratios is given in the results section. To effectively capture high-level spatiotemporal details within these remaining tokens, we utilize the ViT backbone and incorporate joint space-time attention.

This allows all pairs of tokens to interact with each other in the multi-head self-attention layer, enhancing the model's ability to understand complex spatiotemporal relationships.

7.4.2.3 Embedding Generator:

It collects the individual embeddings of all the processed patch tubes and uses a feedforward network to give the joint representation of the entire SITS. It consists of two linear layers.

7.4.2.4 Decoder:

We pre-train SaTran for two tasks viz. reconstruction and classification (refer section 4) using different decoders one for each task.

The decoder for the reconstruction task consists of trans-convolutional layers to perform the upsampling operations and construct the middle frame. Batch normalization is applied to ensure stability and efficiency in training. The decoder is trained using the mean squared error loss function for the model to generate the middle frame reconstructed images.

The decoder for the classification task consists of a simple feedforward network with two linear layers and a sigmoid activation for the output. We use binary cross-entropy loss as the loss function.

The decoders are used only in pre-training and are removed during fine-tuning.

7.5 Pre-training of SaTran

We optimized SaTran for two objectives in pre-training, given as follows:

1. **Reconstruction task:** We pre-train the SatTran using the middle frame reconstruction pretext task. The embeddings obtained from SatTran are passed to an additional decoder to construct the satellite image at the middle timestamp of the time series. The model learns to capture temporal dependencies and patterns inherent in the dynamic changes on Earth's surface. This pre-training task ensures that the model acquires a robust understanding of the temporal and spatial dynamics of SITS.

2. **Classification task:** The second pre-training task is a binary classification task to decide whether the given SITS is ordered or shuffled. SatTran learns whether the given input is temporally in the correct order or not. The model is given as input the original SITS and the shuffled time series by randomly permuting the original time series. The loss function used is binary cross-entropy loss.

We separately pre-trained SatTran on two SITS datasets viz. MODIS (a moderate spatial resolution satellite) and Landsat-8 (high spatial resolution satellite) image time series, and named them SatTran-M and SatTran-L, respectively. This is due to the properties of the satellite data. The characteristics including height, width, bands, and time stamps (i.e. number of frames) are different for different satellites due to the different spatial, temporal, and spectral resolutions. Thus, the model pre-trained for one satellite data cannot be used for the other satellite data. This also requires optimization for suitable size of patch tubes and tiny tubelets for two satellite data. The details are given in the experimentation section 7.7.1.

We have also pre-trained VideoMAE for the reconstruction task in the same way as in the original paper for MODIS data for different masking ratios. The impact of different masking ratios by VideoMAE is given in the results section. VideoMAE did not work for Landsat-8 data in its original resolution because of the huge size of the images. So we pretrained VideoMAE-R, the resized version of VideoMAE given in section 7.6.

Pre-training data details: SatTran is trained for 600 US counties on 7 years (2014-2020) of data. The length of time series taken in pre-training is 13 timestamps for MODIS and 7 timestamps for Landsat-8. The total number of instances used for pre-training is 1M for each satellite data and the size of the dataset is approx. 100GB for MODIS and 900GB for Landsat-8.

Fine-Tuning of SaTran: We have considered six earth observation applications as downstream tasks for testing SaTran. The applications include prediction of crop yield (CYP), snow cover (SCP), solar energy (SEP), soil moisture (SMP), cloud cover (CCP), and classification of land cover (LCP). The data used consists of SITS from two satellites

MODIS and Landsat-8. We have taken around 2000 counties from different states of US for all applications considered. The ground truth data is specific to the application.

7.6 Models for Comparison

We compare SaTran with existing models including two RGB video transformers (VideoMAE [156], and ViViT [34]) and two SITS models - SITSFormer [26] and TSViT [24] developed for classification tasks. VideoMAE is pretrained on the reconstruction task, whereas ViViT is not pretrained and is developed for the classification task. We adopted all four models for both prediction and classification downstream tasks. Also, none of these models is able to process Landsat-8 image time series at its original size and gives out-of-memory (OOM) error. We used two modification techniques in order to use these models for Landsat-8 data:

1. **Resize SITS:** We resize the Landsat-8 image time series to one-fourth along the spatial dimension keeping spectral and time dimensions the same. If the original SITS is of size $B \times T \times H \times W$, the resized SITS is of size $B \times T \times \frac{H}{4} \times \frac{W}{4}$. The variants of the models mentioned above are represented by "$*$-R" in the results section e.g., VideoMAE is represented as VideoMAE-R.

2. **Segmentation:** We segment the original Landsat-8 image time series into 16 segments along the spatial dimension without tampering other two dimensions and each segment is of size $B \times T \times \frac{H}{4} \times \frac{W}{4}$. The existing models are applied to each segment and their embeddings are then concatenated together and passed through a feed-forward network to get the final predictions. This variant is represented by "$*$-S" e.g., VideoMAE is represented as VideoMAE-S.

The details of the models used for comparison are as follows:

1. **VideoMAE [156]:** VideoMAE divides the RGB videos into tubelets and uses a high masking ratio (95%) to handle temporal redundancy in the videos. However, VideoMAE pre-trained on RGB videos is not suitable for SITS data. Thus, for

a fair comparison, we tried to pre-train VideoMAE for the Landsat-8 image time series. We are not able to do so for Landsat-8 data in its original fine resolution due to huge size of images and the GPU was out of memory even after reducing batch size to 2. Thus, we pretrain the *resize* variant of the model mentioned in section 6.1 using 90% and 75% masking ratio and used this variant for all the downstream tasks.

2. **ViViT [34]:** ViViT[34] is originally developed for RGB videos. Image time series of dimension $T \times H \times W$ is converted into multiple tubelets of dimension $z \times h \times w$ and tokens are extracted from all three dimensions. These tubelet embeddings efficiently capture spatiotemporal information by representing non-overlapping spatiotemporal patches. We adapted it for SITS data and used the factorized encoder version of the transformer.

3. **SITSFormer [26]:** SITSFormer uses the neighboring pixels and then predict the label of the center pixel using the BERT model. SITSFormer is originally designed for classification tasks. We have adapted it for prediction tasks by replacing the classification head with two linear layers and an output layer for prediction.

4. **TSViT [24]:** TSViT is also designed for classification tasks. It splits a SITS instance into non-overlapping patches in space and time which are tokenized and processed by a factorized temporal-spatial encoder. It uses class-specific *cls* tokens as inductive bias to improve the model prformance. We have adapted it for prediction tasks by replacing the classification head with prediction head consisting of two linear layers and an output layer for prediction.

7.7 Experiments and Evaluation Metric

7.7.1 Experimental Setup

All the experiments are performed using Pytorch 1.11.0 and CUDA 11.7 on an A100 GPU server having 80GB GPU RAM. SatTran is pre-trained and fine-tuned using Adam

optimizer for 35 and 15 epochs, respectively with a batch size of 8. The learning rate η for the optimizer is decided while tuning hyper-parameters for each model and application. For CYP, CCP, SMP, SEP, SCP, η= [0.0003, 0.001, 0.001, 0.0001, 0.00001], respectively. The size of patch tubes for MODIS is 10×10 and for Landsat-8, it is 64×64. The size of tiny tubelets for MODIS is $5 \times 5 \times 2$ and for Landsat-8 it is $16 \times 16 \times 2$.

We tested SatTran for two years (2019 and 2020) for all the applications. All the predictions are done at timestamp *t*+1, taking 1 to *t* as input TS data. If prediction needs to be done yearly, we averaged the predicted output for the years 2019 and 2020. If the application requires prediction at a monthly level (for SMP and SCP), the average is taken for 24 predictions in two years, and so on.

7.7.2 Evaluation Metrics

We use Root Mean Square Error (RMSE) for prediction problems and accuracy and F1-score for classification. RMSE calculates the root mean squared magnitude of errors, assigning more weight to larger errors.

7.8 Results and Discussion

Preliminary experiments: Pre-training using MODIS data

We present preliminary experiments to pre-train SaTran and the existing model VideoMAE for the reconstruction task using MODIS data. The results are captured for different masking ratios. Table 7.2 presents the GPU memory and the time required for both models using batch size 8 for training. It can be observed from the table that SaTran takes approximately half of the GPU memory than that of VideoMAE. Though VideoMAE needs less time to execute for one epoch in comparison to SaTran, SaTran converges faster due to its attention mechanism of handling spatiotemporal redundancy and distributed approach of applying VideoMAE to patch tubes. Thus the total time taken by SaTran is much less than VideoMAE. Moreover, the GPU memory requirements of VideoMAE increase exponentially by reducing the masking ratio. The RMSE obtained by SaTran is lower than that of VideoMAE in all the cases.

Table 7.2: Memory and Time requirements for SaTran and VideoMAE in pretraining on SITS for Reconstruction task using batch size 8. VideoMAE-R uses reduced size of Landsat-8 SITS.

MODIS Data								
	GPU memory		Time per epoch (in hours)		Total time (in hours)		RMSE	
Masking Ratio	VideoMAE	SaTran	VideoMAE	SaTran	VideoMAE	SaTran	VideoMAE	SaTran
90	30GB	20GB	5.815	5.833	232.612	174.999	0.2052	0.1879
75	39GB	21GB	6.293	6.645	251.724	193.500	0.1984	0.1856
60	49GB	22GB	6.544	6.450	261.776	199.374	0.1902	0.1848
Landsat-8 Data								
Masking Ratio	VideoMAE-R	SaTran	VideoMAE-R	SaTran	VideoMAE-R	SaTran	VideoMAE	SaTran
90	52GB	60GB	5.925	7.259	237.000	217.794	0.2622	0.1927
75	58GB	65GB	6.715	7.864	268.600	235.926	0.2612	0.1901
60	64GB	72GB	6.920	8.764	276.800	262.935	0.2606	0.1893

It can also be observed from the table that though the best results are obtained for 60% masking for both models, the change from 90% to 75% is more than the change from 75% to 60%. However, the computational requirements for 60% are more than 75%. Thus we recommend 75% masking of tiny tubelets this is unlike what is recommended by VideoMAE (95%) for RGB videos. VideoMAE suggests high masking ratio to avoid data leaks. In our case data leaks are avoided by not processing the redundant patches and 75% masking is sufficient to avoid data leaks in non-redundant patches. We also conducted experiments for batch size 16 where VideoMAE gives the out-of-memory error when the masking ratio is reduced to 60%. The results are given in ablation study subsection.

Pre-training using Landsat-8 data

The memory and time requirements for the models pretrained using Landsat-8 data are given in Table 7.2. We use VideoMAE-R variant for Landsat-8 data because Landsat-8 is a high spatial resolution satellite and its image time series is huge and cannot be processed using VideoMAE even for the masking ratio of 90% on the systems with the specifications mentioned in section 6. Whereas, SaTran is successfully able to process Landsat-8 image time series in its original resolution and also requires less time as compared to VideoMAE but the GPU memory used is slightly more than that for VideoMAE-R.

Comparison of SaTran with existing models:

We compare SaTran with existing models (given in section 6.1) for various downstream tasks using Landsat-8 data. Table 7.3 represents the RMSE obtained by different models for various downstream tasks. It can be observed from the table that none of the existing models were able to process the Landsat-8 time series with its original dimensions due to its large size. All the existing models suffer from out-of-memory (OOM) error. The *resize* and *segmentation* variants of models are able to process Landsat-8 SITS, but their performance is inferior to that of SaTran. SITSFormer and TSViT perform better for classification problems as they were originally developed for similar classification problems and we adapted them for prediction tasks.

Table 7.3: Comparison of SaTran with existing transformer and SITS models for Landsat-8 data for different earth observation applications for prediction and classification using RMSE and F1-score, respectively. VideoMAE, ViViT, SITSFormer, and TSViT, all the models through OOM error while using original resolution of Landsat-8 image time series

	Prediction Problems (RMSE)					Classification (LCP)		
Model	CYP	SEP	SCP	SMP	CCP	Accuracy	Precision	F1-score
VideoMAE-R (90%) [156]	9.2038	6.7957	22.2991	7.5549	18.6554	0.6823	0.4589	0.4397
VideoMAE-R (75%) [156]	7.3457	6.3511	19.1947	6.7957	18.8426	0.7101	0.4980	0.4605
SITSFormer-R [26]	6.9860	6.8325	19.8670	5.9716	16.2534	0.8053	0.5763	0.5874
ViViT-R [34]	8.7552	6.4307	22.0058	6.3987	18.3996	0.6915	0.5694	0.5201
TSViT-R [24]	9.3365	7.1937	19.8471	4.7591	17.8773	0.8368	0.6923	0.6186
VideoMAE-S (90%) [156]	8.6085	6.1708	21.4295	7.1644	18.3846	0.7462	0.5031	0.4703
VideoMAE-S (75%) [156]	7.3457	5.9611	17.5016	6.4978	17.7153	0.7551	0.5112	0.4957
SITSFormer-S [26]	6.6661	6.4081	19.7654	5.4121	15.3935	0.8289	0.6041	0.6153
ViViT-S [34]	7.1626	6.0159	21.6651	6.0159	16.7468	0.7263	0.6198	0.5562
TSViT-S [24]	8.3497	6.1563	17.2997	4.4878	17.0686	0.8506	0.7212	0.6415
SaTran (75%) (our)	**5.5584**	**5.0193**	**15.3112**	**3.0775**	**12.0716**	**0.9486**	**0.8172**	**0.6967**

Table 7.4: Comparison of SaTran with competitive models for various downstream applications using MODIS data

Model	CYP	SEP	SCP	SMP	CCP
VideoMAE (90%) [156]	8.1661	6.4693	21.1754	6.5409	14.7517
VideoMAE (75%) [156]	7.9826	6.3983	20.7469	5.1252	14.7325
SITSFormer[26]	8.2761	7.5721	19.9875	6.1574	19.2559
ViViT [34]	9.2437	6.7068	20.0251	6.4843	17.2785
TSViT [24]	9.1051	6.2486	17.4759	4.9115	18.6722
SaTran (75%) (our)	**6.5825**	**5.7992**	**16.8472**	**3.3842**	**12.3456**

It can also be observed that the error obtained by the *resize* variant of the models is larger than their corresponding *segmentation* variants. This is because the spatial resolution of SITS is degraded. In a few cases, the performance of *resize* variants degrades even from MODIS data (see Table 7.4) because the temporal resolution of Landsat-8 is already coarser than MODIS and due to resizing the spatial resolution is also compromised, thus leading to further loss of information. On the other hand, the proposed model SaTran not only successfully processes the Landsat-8 image time series at its original resolution but also outperforms both the variants of all existing models for all the tasks.

Memory and time requirements of the Models

The GPU memory, time, and number of training parameters are given in Table 7.5. It can be observed from the table that only ViViT has lesser number of training parameters than that of SaTran, and all other models (both variants) need more training parameters than SaTran. Also, the training time of SaTran is comparable to *resize* variant of existing models and it is lesser than the *segmentation* variants which use the same resolution of SITS as that used in SaTran. It can also be observed that the GPU memory requirements of SaTran are also comparable to the competing models.

It is evident from Table 7.3 and 7.5 that SaTran outperforms all the existing models and has reasonable time and space requirements. None of the baselines either processes the Landsat image time series at coarser spatial resolution (*resize* variant) or by segmenting in the spatial dimension performs well in solving the earth observation applications. Thus, this establishes the requirement of SaTran which can efficiently process large size

Table 7.5: Comparison of SaTran with competitive models for memory and time requirements for CYP

Model	# training parameters	Training time (in hours)	GPU memory
VideoMAE-R (90%) [156]	431M	14.5	48GB
VideoMAE-R (75%) [156]	431M	16.0	54GB
SITSFormer-R [26]	500M	18.0	42GB
ViViT-R [34]	86M	9.16	52GB
TSViT-R [24]	360M	12.5	56GB
VideoMAE-S (90%) [156]	451M	27.5	67GB
VideoMAE-S(75%) [156]	451M	29.1	72GB
SITSFormer-S [26]	553M	48.0	58GB
ViViT-S [34]	92M	15.3	72GB
TSViT-S [24]	490M	22.2	66GB
SaTran (75%)	311M	16.4	64GB

SITS to give learned representations and can be used for various applications.

We have also conducted experiments for MODIS image time series for fair comparison because the existing models are able to process the MODIS data in its original spatial resolution. The results are given in Table 7.4. The results show that SaTran outperforms all the existing models.

7.8.1 Ablation Study

We performed ablation study for deciding traversal fraction (1/x) of SITS, batch size, masking ratio for tiny tubelets, and pretraining tasks. The details are given in Appendix D. We fixed masking ratio for the experiments at 75%. We concluded empirically that optimal traversal for MODIS x=4 (refer Figure 7.3) and for Landsat x=3 (refer Figure 7.4), respectively. We used these selected values of all parameters in all the experiments.

Results of MODIS for Batch size 16:

We also conducted experiments for different masking ratio and batch size while training. When batch size is increased to 16, VideoMAE is not able to process MODIS time series also if the masking ratio is less than 75%. However, SaTran performs well even at a lower masking ratio as well. Thus, for fair comparison, we performed all results at 8 batch size. Also, this shows that SaTran is an efficient model in terms of computation requirements

Table 7.6: Memory and Time required for MODIS with batch size 16

	GPU memory		Time per epoch (in hours)		Total time (in hours)		RMSE	
Masking	VideoMAE	SaTran	VideoMAE	SaTran	VideoMAE	SaTran	VideoMAE	SaTran
90	54 GB	30GB	3.5472	3.6111	177.3600	126.3891	0.2081	0.1874
75	70GB	31GB	3.6694	3.7222	183.4700	130.2775	0.1975	0.1861
60	OOM	32GB	OOM	4.0292	OOM	141.0220	OOM	0.1818

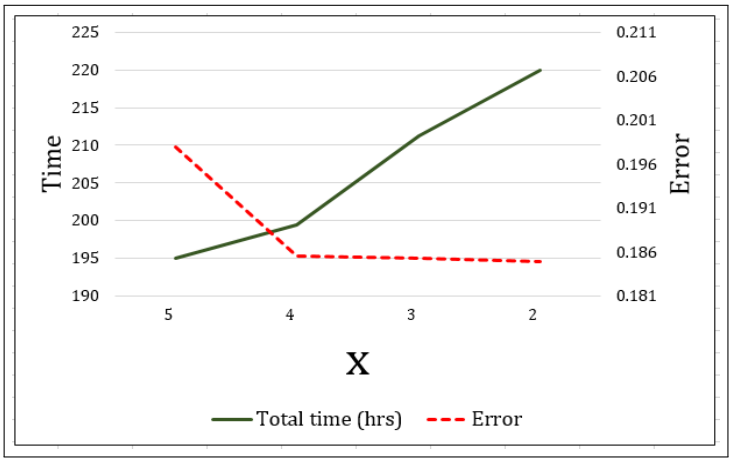

Figure 7.3: Deciding x for (1/x)th traversal of SITS: MODIS Data

in comparison to VideoMAE for SITS analytics. The results are given in Table 7.6.

Selecting optimal x for partial traversal of SITS

Figure 7.3 and 7.4 shows the computation time required and error curve for the reconstruction task by varying x as 5,4,3, and 2 for MODIS and Landsat-8 time series, respectively. The masking ratio is fixed at 75%. It can be observed from the figure that, there is an improvement of $\approx 6\%$ in the performance of the SaTran-M when we traverse $1/x$ of the time series x changing x from 5 to 4, and the error did not change much after that. Similarly, for SaTran-L, the maximum improvement is observed when x changes from 4 to 3. The computation time required from x = 5 to 4 for SaTran-M and from x = 4 to 3 does not increase much, but it increases linearly after that. The results for the downstream tasks are given for the optimal traversal in both cases (For MODIS x=4 and for Landsat x=3).

Impact of different masking ratios on downstream applications

To analyze the impact of different masking ratios of tiny tubelets, we performed exper-

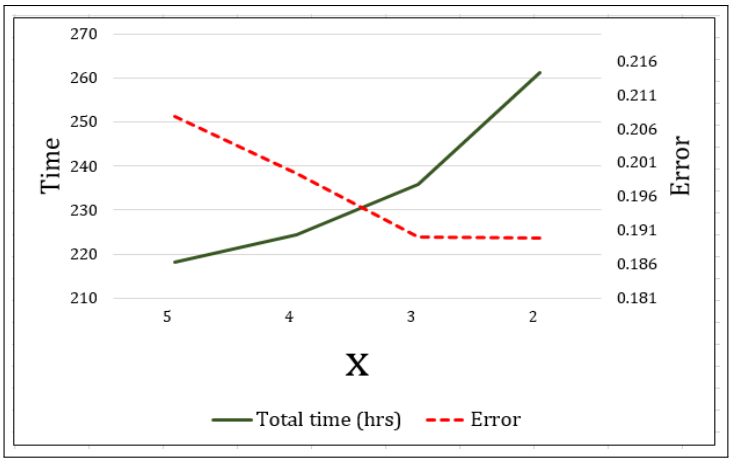

Figure 7.4: Deciding x for (1/x)th traversal of SITS: Landsat-8 Data

Table 7.7: Impact of masking ratios on performance of SaTran and VideoMAE (pre-trained on Reconstruction Task using MODIS data) for various applications (RMSE)

Application	**90% masking**		**75% masking**		**60% masking**	
	VideoMAE	**SaTran**	**VideoMAE**	**SaTran**	**VideoMAE**	**SaTran**
CYP	8.1661	7.7240	7.9826	7.3894	7.5549	7.3862
SCP	21.175	18.175	20.746	18.115	20.449	18.057
SEP	6.4693	6.3361	6.3983	6.1997	6.2302	6.1806
CCP	14.751	14.342	14.732	13.810	14.324	13.764
SMP	6.5409	3.8088	5.1252	3.5338	4.8953	3.5131

iments for prediction downstream applications using MODIS data. The table 7.7 gives RMSE for various applications. It can be observed from Table that SaTran outperforms VideoMAE for all the applications for all masking ratios. For example, when using 75% masking, RMSE reduces by approximately 8%, 14.5%, 3%, 6.6%, and 45% for soybean, snow cover, solar energy, cloud cover, and soil moisture prediction, respectively.

Impact of pretraining tasks on SaTran

Table 7.8 gives RMSE values and shows the impact of training SaTran for different pre-training tasks. The results in table are presented with 1/4 traversal of MODIS image time series and 75% masking of the tubelets. It can be observed that RMSE reduces by a significant margin for all the applications when SaTran is pre-trained further using classification task. In case of MODIS data, maximum improvement is seen in crop yield prediction with approximately 12% reduction in RMSE, followed by 11% in cloud cover

Table 7.8: Impact of different pretraining tasks on performance of SaTran for various downstream applications

MODIS Data			
Application	Pre-trained on reconstruction task	Pre-trained on reconstruction + classification tasks	% improvement
CYP	7.3894	6.5825	12.25
SCP	18.115	16.847	7.520
SEP	6.1997	5.7992	6.900
CCP	13.810	12.345	11.86
SMP	3.5338	3.3842	4.420
Landsat-8 Data			
Application	Pre-trained on reconstruction task	Pre-trained on reconstruction + classification tasks	% improvement
CYP	5.9046	5.5584	6.220
SCP	15.641	15.311	2.150
SEP	5.2523	5.0193	4.640
CCP	12.195	12.071	1.020
SMP	3.4844	3.0775	13.22

prediction. The minimum improvement observed is 4% for soil moisture prediction.

Similarly, Table 7.8 presents the RMSE obtained by SaTran when pre-trained only for the reconstruction task and both reconstruction and classification tasks using high spatial resolution Landsat-8 image time series. The results in table are presented with 1/3 traversal of SITS and 75% masking of the tubelets for Landsat-8. It can be observed from the table that RMSE reduces when SaTran is pre-trained for both tasks. The maximum improvement is observed in soil moisture prediction with a reduction in RMSE by 13%.

7.9 Main Contributions

The main contributions of the chapter are listed below:

1. We introduce SaTran, a novel, cost and time-efficient transformer for large-size SITS to learn their generic representation for earth observation tasks.

2. SaTran has a twofold redundancy handling mechanism which ignores 1) spatiotemporal redundant patch tubes and 2)temporally redundant spans in non-redundant patch tubes. Moreover, SaTran uses a distributed processing approach to apply

VideoMAE on non redundant patch tubes. The redundancy handling and distributed approach collectively result in reduced memory and time requirements.

3. We have done extensive experimentation and compared our model with existing video vision models and SITS models. The results show that SaTran outperforms the existing models for various downstream earth observation applications like crop yield prediction, snow cover prediction, land cover classification, etc.

7.10 Summary

We have presented a two-fold data redundancy handling, self-supervised learning method SatTran for SITS transformer pretraining. We pretrained SatTran for two pretext tasks- Reconstruction and classification and fine-tuned for six earth observation applications including prediction of crop yield, soil moisture, solar energy, snow cover, cloud cover, and classification of land cover. SaTran introduces two novel designs - automatic patch tube selection and a distributed approach of applying tube masking on tiny tubelets. SatTran reduces the memory requirements by approximately a factor of 2. Experimental results show that due to short temporal redundancy, it is not recommended to have a very high masking ratio to achieve better results. Our experiments also demonstrate that the time taken by SatTran increases sublinearly with the increase in image size e.g., we observed an increase of 18% in processing time for 900GB of Landsat-8 data in comparison to 100GB of moderate resolution SITS (MODIS) data. The proposed transformer model outperforms existing video models and SITS transformers for all the downstream earth observation applications considered.

Chapter 8

Conclusions and Future Directions

8.1 Conclusion

In this thesis, we have explored the diverse realm of data used in earth observation applications, distinguishing between conventional proximally sensed data and satellite data. We have worked with different modalities of satellite data viz. histograms, images, and spectral reflectance indices. In addition to this, we have also included other modalities like meteorological, and soil data wherever applicable.

We have given two crop yield prediction models for various crops spanning different regions in India and US. YPN is based on proximally sensed data and CYN incorporated histograms obtained from satellites. We innovatively modeled CYP as a spatiotemporal problem, incorporating effective design decisions such as spatial clustering, and padded crop cycles. Additionally, we introduced attribute selection and depth selection modules to enhance prediction accuracy, making recommendations crucial for optimizing any crop yield prediction system. Our experiments underline the importance of modeling depth-variant soil information, significantly improving model performance. We compared CYN with existing models for their generalizability capability. The increase in RMSE for corn using models trained on both corn and soybean is 16.37%, 9.77%, 5.57%, and 0.68% for

CNN, CNN+GP, CNN+LSTM, and CYN, respectively.

Furthermore, we addressed challenges associated with historic data scarcity in satellite imagery, particularly in context of high spatial resolution datasets like Landsat-8 and Sentinel-2. Our findings highlighted the efficacy of data augmentation techniques in improving yield prediction accuracy, especially when utilizing high-resolution satellite data.

PatchNet, a model for processing high spatial resolution satellite image time series, demonstrated its superiority over existing models, leveraging spatial information and temporal dynamics for finer-grained analysis in earth observation applications. The experimentation shows that PatchNet outperforms the existing models working with histogram time series of satellite data. By preserving spatial information and temporal dynamics, using image time series provides a richer representation of changes in the Earth's surface.

The democratization of satellite image technology remains hindered by processing constraints and the limited availability of high-resolution data. Satellite image technology holds immense promise for global coverage and data provision, yet challenges persist, including the trade-off between spatial and temporal resolution. We proposed fusion techniques—LSFuseNet and FuSITSNet that effectively address this trade-off by leveraging fusion modules and feature alignment mechanisms. These techniques offer a robust solution for earth observation tasks, by seamlessly integrating high spatial and temporal features without increasing data volume, thus overcoming a significant hurdle in the field.

Moreover, we contributed to the optimization of spectral reflectance index selection, introducing the innovative concept of SRI images and developing the SpInN model for automated recommendation of relevant SRIs. SpInN, leveraging ViViT and BERT architectures, emerged as a state-of-the-art solution for various Earth observation applications, offering automation and efficiency in spectral reflectance index recommendation.

Lastly, we proposed a transformer SaTran for SITS. It uses an automatic patch tube selection mechanism which ignores spatiotemporally redundant patches and exploits the spatial correlation between pixels by processing of patch tubes and handling of their temporal redundancy using tube masking using RGB transformer VideoMAE. This two-fold handling of redundancy enables space and time-efficient processing of SITS.

Table 8.1: Summary of models proposed in the thesis

Model	Description	Experiments done on
YieldPredictNet	Introduces spatiotemporal modeling of crop yield prediction and depth level selection modelling for soil	Conventionally collected data (meteorological and soil)
CropYieldNet	For spatiotemporal modeling of crop yield prediction using satellite data and incorporate data from multiple modalities – satellite, meteorological (time series), and soil (static).	Histogram time series obtained from MODIS, Landsat-8, and Sentinel-2
PatchNet	For efficient representation learning of image time series of satellite with high spatial resolution	Image time series of Landsat-8
LSFuseNet	Fuses histogram time series of two satellites of varying spatial, temporal and spectral resolution	Histogram time series of Landsat-8 and Sentinel-2
FuSITSNet	Feature level fusion model for image time series of two satellites each having either high spatial resolution or high temporal resolution	Image time series of MODIS and Landsat-8
SpInN	Recommends spectral reflectance indices for various earth observation applications and solve these problems using selected SRIs	SRI image time series obtained from MODIS data.
SaTran	Foundation model for processing large size SITS where existing vision/video models shortfall	MODIS and Landsat-8 image time series

The meticulous experimentation, comparative analysis, ablation study, and innovative techniques proposed in the thesis attempt to significantly advance the area of Artificial Intelligence for Earth Observation (AI4EO). The thesis endeavors to further the research at the intersection of deep learning and satellite image technologies. All the proposed solutions can be fine-tuned for application to different regions of the world and for emerging earth observation applications. The summary of models proposed in the thesis is given in Table 8.1

8.1.1 Limitations

Due to lack of availability of high spatial resolution satellites, we could not test on finer resolution. Most of them are not available to public free of cost.

8.2 Future Directions

1. The proposed models can be upgraded by incorporating:
    a. physics-informed characteristics
    b. knowledge-enhanced models.
2. The hyperspectral imagery can be used for enhancing our models and adapting for applications like crop classification, soil characteristic monitoring, crop disease detection, green-house gas emission, etc.
3. Some applications like crop yield prediction and soil moisture can be extended to solve at smaller granularity like farm level.
4. Models can be enhanced for Crop yield prediction for the next crop cycle prediction.

Appendix A

Landsat-8 Data Preprocessing

Since the satellite data is captured as raw multispectral images. It needs data pre-processed before it can be used for the end task. The data preparation steps are given below:

A.0.0.1 Bits Precision

By default the Landsat-8 images have float values at every pixel for all the reflectance bands. Images in float values require 32 bits to store a single pixel. The number of pixels in a Landsat-8 image on average is 2000×2000. Thus it requires a large storage space for an image to store. To work with the mentioned spatiotemporal applications, we need historical data for all the locations, thus storing so many images of such huge size is difficult. To save storage space we used the bits precision compression technique explained in Chapter 2 section 2.5.1.1.

A.0.0.2 Time series length in each application

The length of the time series taken for each application is different and also depends on the satellite. For CYP, we have used a padded crop cycle for each crop by padding 2 time stamps on either side of a crop cycle. Taking a padded crop cycle for CYP captures any anomalies or sudden changes in climatic conditions before sowing a crop which can have a substantial effect on crop growth. Additionally, padding the extra time stamps to the crop cycle handles the discrepancy in sowing and harvesting dates across locations.

Table A.1: Length of time series used

Satellite	CYP		SCP	SEP
	Corn	Soybean		
MODIS	36	32	12	24
Landsat-8	18	16	6	12

For snow cover and solar energy prediction we have taken the time series from the last three months to predict snow cover for the next month and solar energy for next fortnight. These applications depend on climatic conditions which change over a few days. The exact time series length for all three applications in the case of the two satellites is given in Table A.1.

Bibliography

[1] "Landsat 1 — U.S. Geological Survey — usgs.gov." https://www.usgs.gov/landsat-missions/landsat-1. [Accessed 18-01-2024].

[2] h. Market Research Future, "Satellite Data Services Market Size & Overview, Industry, Growth — marketresearchfuture.com." https://www.marketresearchfuture.com/reports/satellite-data-services-market-8562. [Accessed 27-01-2024].

[3] B. Ferreira, M. Iten, and R. G. Silva, "Monitoring sustainable development by means of earth observation data and machine learning: A review," *Environmental Sciences Europe*, vol. 32, no. 1, pp. 1–17, 2020.

[4] P. Micheuz, "Approaches to artificial intelligence as a subject in school education," in *Empowering Teaching for Digital Equity and Agency: IFIP TC 3 Open Conference on Computers in Education, OCCE 2020, Mumbai, India, January 6–8, 2020, Proceedings*, pp. 3–13, Springer, 2020.

[5] M. A. Nielsen, *Neural networks and deep learning*, vol. 25. Determination press San Francisco, CA, USA, 2015.

[6] "Neural network types — mriquestions.com." https://mriquestions.com/deep-network-types.html. [Accessed 21-02-2024].

[7] J. Dancker, "A Brief Introduction to Recurrent Neural Networks — towardsdatascience.com." https://towardsdatascience.com/

a-brief-introduction-to-recurrent-neural-networks-638f64a61ff4. [Accessed 07-02-2024].

[8] "Architectures &x2014; ML Glossary documentation — ml-cheatsheet.readthedocs.io." https://ml-cheatsheet.readthedocs.io/en/latest/architectures.html. [Accessed 07-02-2024].

[9] "Application of remote sensing to study forest fires — sciencedirect.com." https://www.sciencedirect.com/science/article/abs/pii/B9780323992626000158. [Accessed 17-01-2024].

[10] NASA, "Modis." https://lpdaac.usgs.gov/data/get-started-data/collection-overview/, 2015. Accessed: 2022-12-06.

[11] "LP DAAC - MODIS Overview — lpdaac.usgs.gov." https://lpdaac.usgs.gov/data/get-started-data/collection-overview/missions/modis-overview/. [Accessed 19-01-2024].

[12] "Explore Copernicus satellite missions - Sentinel Online — sentinels.copernicus.eu." https://sentinels.copernicus.eu/web/sentinel/home. [Accessed 19-01-2024].

[13] "Satellite Imagery Analytics — Planet — planet.com." https://www.planet.com/products/planet-imagery/?utm_campaign=gisint&utm_content=pros-leads-gisresponsivesearch-0923&utm_source=google&utm_medium=paid-search&gad_source=1&gclid=CjwKCAiA98WrBhAYEiwA2WvhOu0dmmXan6MTV6a0Y93JAR6UC-5JhPOCQzY_KjJtm2Lgf_IOIma5lBoCM2kQAvD_BwE&restored=1701943276547. [Accessed 19-01-2024].

[14] https://crisp.nus.edu.sg/~research/tutorial/spot.htm. [Accessed 19-01-2024].

[15] Y. Quan, Y. Tong, W. Feng, G. Dauphin, W. Huang, and M. Xing, "A novel image fusion method of multi-spectral and sar images for land cover classification," *Remote Sensing*, vol. 12, no. 22, p. 3801, 2020.

[16] S. Pan, H. Guan, Y. Chen, Y. Yu, W. N. Gonçalves, J. M. Junior, and J. Li, "Land-cover classification of multispectral lidar data using cnn with optimized hyper-parameters," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 241–254, 2020.

[17] Y. Yu, T. Jiang, J. Gao, H. Guan, D. Li, S. Gao, E. Tang, W. Wang, P. Tang, and J. Li, "Capvit: Cross-context capsule vision transformers for land cover classification with airborne multispectral lidar data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 111, p. 102837, 2022.

[18] C. Qiu, L. Mou, M. Schmitt, and X. X. Zhu, "Fusing multiseasonal sentinel-2 imagery for urban land cover classification with multibranch residual convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 10, pp. 1787–1791, 2020.

[19] X. Zhang, L. Liu, X. Chen, Y. Gao, S. Xie, and J. Mi, "Glc_fcs30: Global land-cover product with fine classification system at 30 m using time-series landsat imagery," *Earth System Science Data Discussions*, pp. 1–31, 2020.

[20] S. Son and J. Kim, "Land cover classification map of northeast asia using goci data," *Korean Journal of Remote Sensing*, vol. 35, no. 1, pp. 83–92, 2019.

[21] Y. Gupta, N. Goyal, V. J. Varghese, and P. Goyal, "Utilizing modis fire mask for predicting forest fires using landsat-9/8 and meteorological data," in *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10, IEEE, 2023.

[22] "Satellite Imaging Market Size, Share, Growth — Overview, 2030 — fortunebusinessinsights.com." `https://www.fortunebusinessinsights.com/satellite-imaging-market-103372`. [Accessed 27-01-2024].

[23] A. Debien, M. Casaburi, G. Milcinski, and M. Maranesi, "Esa's ai4eo initiative: Bridging the gap between the ai & earth observation communities," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 251–253, IEEE, 2021.

[24] M. Tarasiou, E. Chavez, and S. Zafeiriou, "Vits for sits: Vision transformers for satellite image time series," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10418–10428, 2023.

[25] Y. Yuan and L. Lin, "Self-supervised pretraining of transformers for satellite image time series classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 474–487, 2020.

[26] Y. Yuan, L. Lin, Q. Liu, R. Hang, and Z.-G. Zhou, "Sits-former: A pre-trained spatio-spectral-temporal representation model for sentinel-2 time series classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 106, p. 102651, 2022.

[27] F. Gao, J. Masek, M. Schwaller, and F. Hall, "On the blending of the landsat and modis surface reflectance: Predicting daily landsat surface reflectance," *IEEE Transactions on Geoscience and Remote sensing*, vol. 44, no. 8, pp. 2207–2218, 2006.

[28] L. Zhang and Q. Weng, "Annual dynamics of impervious surface in the pearl river delta, china, from 1988 to 2013, using time series landsat imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 113, pp. 86–96, 2016.

[29] S. Li, J. T. Kwok, and Y. Wang, "Using the discrete wavelet frame transform to

merge landsat tm and spot panchromatic images," *Information Fusion*, vol. 3, no. 1, pp. 17–23, 2002.

[30] S. Li and B. Yang, "A new pan-sharpening method using a compressed sensing technique," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 2, pp. 738–746, 2010.

[31] H. K. Zhang and D. P. Roy, "Computationally inexpensive landsat 8 operational land imager (oli) pansharpening," *Remote sensing*, vol. 8, no. 3, p. 180, 2016.

[32] R. A. Schwalbert, T. Amado, G. Corassa, L. P. Pott, P. V. Prasad, and I. A. Ciampitti, "Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern brazil," *Agricultural and Forest Meteorology*, vol. 284, p. 107886, 2020.

[33] M. L. Hunt, G. A. Blackburn, L. Carrasco, J. W. Redhead, and C. S. Rowland, "High resolution wheat yield mapping using sentinel-2," *Remote Sensing of Environment*, vol. 233, p. 111410, 2019.

[34] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6836–6846, 2021.

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[36] N. Kumar, *Implementation of the NC-94 hybrid storage prototype on a binary version of CanStoreX*. Iowa State University, 2008.

[37] Google, "Google Earth Engine — earthengine.google.com." `https://earthengine.google.com/`, 2019. Accessed: 2022-5-5.

[38] "Microsoft Planetary Computer — planetarycomputer.microsoft.com." https://planetarycomputer.microsoft.com/. [Accessed 20-04-2024].

[39] K. Choudhary, V. Pandey, C. Murthy, and M. Poddar, "Synergetic use of optical, microwave and thermal satellite data for non-parametric estimation of wheat grain yield," *The Int. Archives of Photogrammetry, RS and Spatial Information Sciences*, vol. 42, pp. 195–199, 2019.

[40] NASA, "Landsatwebpage." https://www.usgs.gov/faqs/what-are-band-designations-landsat-satellites?qt-news_science_products=0#qt-news_science_products, 2016. Accessed: 2022-2-16.

[41] European Space Agency Signature, "Sentinel webpage." https://www.netiq.com/documentation/sentinel-82/user/data/bookinfo.html, 2017. Accessed: 2023-2-25.

[42] paul, "Interpolation methods — paulbourke.net." http://paulbourke.net/miscellaneous/interpolation/. [Accessed 06-june-2022].

[43] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep gaussian process for crop yield prediction based on remote sensing data," in *Thirty-First AAAI conference on artificial intelligence*, 2017.

[44] J. Sun, Z. Lai, L. Di, Z. Sun, J. Tao, and Y. Shen, "Multilevel deep learning network for county-level corn yield estimation in the us corn belt," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5048–5060, 2020.

[45] B. Verma, R. Prasad, P. K. Srivastava, S. A. Yadav, P. Singh, and R. Singh, "Investigation of optimal vegetation indices for retrieval of leaf chlorophyll and leaf area index using enhanced learning algorithms," *Computers and Electronics in Agriculture*, vol. 192, p. 106581, 2022.

[46] A. Kern, Z. Barcza, H. Marjanović, T. Árendás, N. Fodor, P. Bónis, P. Bognár, and J. Lichtenberger, "Statistical modelling of crop yield in central europe using climate data and remote sensing vegetation indices," *Agricultural and Forest Meteorology*, vol. 260, pp. 300–320, 2018.

[47] "OpenLandMap Soil Organic Carbon Content — developers.google.com." `https://developers.google.com/earth-engine/datasets/catalog/OpenLandMap_SOL_SOL_ORGANIC-CARBON_USDA-6A1C_M_v02`. [Accessed 20-04-2024].

[48] S. Khaki, L. Wang, and S. V. Archontoulis, "A cnn-rnn framework for crop yield prediction," *Frontiers in Plant Science*, vol. 10, p. 1750, 2020.

[49] Y. Everingham, J. Sexton, D. Skocaj, and G. Inman-Bamber, "Accurate prediction of sugarcane yield using a random forest algorithm," *Agronomy for sustainable development*, vol. 36, no. 2, pp. 1–9, 2016.

[50] N. Gandhi and L. J. Armstrong, "A review of application of data mining techniques for decision making in agriculture," in *2nd IC3I*, IEEE, 2016.

[51] P. Perner, *Advances in Data Mining. Applications and Theoretical Aspects: 9th ICDM 2009, Leipzig, Germany, July 20-22, 2009. Proceedings*, vol. 5633. Springer Science & Business Media, 2009.

[52] A. Kumar, N. Kumar, and V. Vats, "Efficient crop yield prediction using machine learning algorithms," *International Research Journal of Engineering and Technology*, vol. 5, no. 06, pp. 3151–3159, 2018.

[53] A. Crane-Droesch, "Machine learning methods for crop yield prediction and climate change impact assessment in agriculture," *Environmental Research Letters*, vol. 13, no. 11, p. 114003, 2018.

[54] P. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction with deep cnn," *Comp. and elect. in agriculture*, vol. 163, p. 104859, 2019.

[55] P. Bose, N. K. Kasabov, L. Bruzzone, and R. N. Hartono, "Spiking neural networks for crop yield estimation based on spatiotemporal analysis of image time series," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 11, pp. 6563–6573, 2016.

[56] J. L. Fernandes, N. F. F. Ebecken, and J. C. D. M. Esquerdo, "Sugarcane yield prediction in brazil using ndvi time series and neural networks ensemble," *International Journal of Remote Sensing*, vol. 38, no. 16, pp. 4631–4644, 2017.

[57] V. Puri, A. Nayyar, and L. Raja, "Agriculture drones: A modern breakthrough in precision agriculture," *J. Stats. & Mgmt. Systems*, vol. 20, no. 4, 2017.

[58] J. Sun, L. Di, Z. Sun, Y. Shen, and Z. Lai, "County-level soybean yield prediction using deep cnn-lstm model," *Sensors*, vol. 19, no. 20, p. 4363, 2019.

[59] U. Verma, H. Piepho, A. Goyal, J. Ogutu, M. Kalubarme, *et al.*, "Role of climatic variables and crop condition term for mustard yield prediction in haryana," *Int J Agric Stat Sci*, vol. 12, pp. 45–51, 2016.

[60] W. Jiu-jiang, W. Nan, S. Hong-zheng, and M. Xiao-yi, "Spatial–temporal variation of climate & its impact on winter wheat production in guanzhong plain, china," *Comp. & Elect. in Agriculture*, vol. 195, p. 106820, 2022.

[61] P. K. Jha, P. Athanasiadis, S. Gualdi, A. Trabucco, V. Mereu, V. Shelia, and G. Hoogenboom, "Using daily data from seasonal forecasts in dynamic crop models for yield prediction: a case study for rice in nepal's terai," *Agricultural and forest meteorology*, vol. 265, pp. 349–358, 2019.

[62] N. Balakrishnan and G. Muthukumarasamy, "Crop production-ensemble machine learning model for prediction," *International Journal of Computer Science and Software Engineering*, vol. 5, no. 7, p. 148, 2016.

[63] S. Sharma and S. Chatterjee, "Corn yield prediction in us midwest using artificial neural networks,"

[64] R. B. Guruprasad, K. Saurav, and S. Randhawa, "Machine learning methodologies for paddy yield estimation in india: a case study," in *IGARSS 2019-2019*, pp. 7254–7257, IEEE, 2019.

[65] S. Khaki and L. Wang, "Crop yield prediction using deep neural networks," *Frontiers in plant science*, vol. 10, p. 621, 2019.

[66] J. Shook, T. Gangopadhyay, L. Wu, B. Ganapathysubramanian, S. Sarkar, and A. K. Singh, "Crop yield prediction integrating genotype and weather variables using deep learning," *Plos one*, vol. 16, no. 6, p. e0252402, 2021.

[67] H. Måløy, S. Windju, S. Bergersen, M. Alsheikh, and K. L. Downing, "Multimodal performers for genomic selection and crop yield prediction," *Smart Agricultural Technology*, vol. 1, p. 100017, 2021.

[68] M. M. Rahman and A. Robson, "Integrating landsat-8 & sentinel-2 time series data for yield prediction of sugarcane at block level," *RS*, vol. 12, no. 8, p. 1313, 2020.

[69] S. Skakun, E. Vermote, B. Franch, J.-C. Roger, N. Kussul, J. Ju, and J. Masek, "Winter wheat yield assessment from landsat 8 and sentinel-2 data: Incorporating surface reflectance, through phenological fitting, into regression yield models," *Remote Sensing*, vol. 11, no. 15, p. 1768, 2019.

[70] L. K. Petersen, "Real-time prediction of crop yields from modis relative vegetation health: A continent-wide analysis of africa," *Remote Sensing*, vol. 10, no. 11, p. 1726, 2018.

[71] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education India, 2016.

[72] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[73] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[74] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[75] J. P. Da Silva, J. Zullo, and L. A. Romani, "A time series mining approach for agricultural area detection," *IEEE Trans. on Big Data*, vol. 6, p. 537–546, 2020.

[76] G. Ruan, X. Li, F. Yuan, D. Cammarano, S. T. Ata-UI-Karim, X. Liu, Y. Tian, Y. Zhu, W. Cao, and Q. Cao, "Improving wheat yield prediction integrating proximal sensing and weather data with machine learning," *Comp. & Elect. in Agriculture*, vol. 195, p. 106852, 2022.

[77] M. S. Divakar, M. S. Elayidom, and R. Rajesh, "Forecasting crop yield with deep learning based ensemble model," *Materials Today: Proceedings*, vol. 58, pp. 256–259, 2022.

[78] P. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction with deep convolutional neural networks," *Computers and electronics in agriculture*, vol. 163, p. 104859, 2019.

[79] S. Khaki, H. Pham, and L. Wang, "Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning," *Scientific Reports*, vol. 11, no. 1, pp. 1–14, 2021.

[80] USDA, "Usda/nass quickstats ad-hoc query tool." `https://quickstats.nass.usda.gov/`, 2010. Accessed: 2022-7-15.

[81] "ICRISAT x2013; Science of discovery to science of delivery — icrisat.org." `https://www.icrisat.org/`. [Accessed 02-may-2022].

[82] J. Pöppelbaum, G. S. Chadha, and A. Schwung, "Contrastive learning based self-supervised time-series analysis," *Applied Soft Computing*, vol. 117, p. 108397, 2022.

[83] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in NIPS*, vol. 33, pp. 18661–18673, 2020.

[84] A. Kaur, P. Goyal, K. Sharma, L. Sharma, and N. Goyal, "A generalized multi-modal deep learning model for early crop yield prediction," in *International Conference on Big Data*, pp. 1272–1279, IEEE, 2022.

[85] T. Sakamoto, "Incorporating environmental variables into a modis-based crop yield estimation method for united states corn and soybeans through the use of a random forest regression algorithm," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 160, pp. 208–228, 2020.

[86] S. Skakun, N. I. Kalecinski, M. G. Brown, D. M. Johnson, E. F. Vermote, J.-C. Roger, and B. Franch, "Assessing within-field corn and soybean yield variability from worldview-3, planet, sentinel-2, and landsat 8 satellite imagery," *Remote Sensing*, vol. 13, no. 5, p. 872, 2021.

[87] Z. Ji, Y. Pan, X. Zhu, D. Zhang, and J. Wang, "A generalized model to predict large-scale crop yields integrating satellite-based vegetation index time series and phenology metrics," *Ecological Indicators*, 2022.

[88] NASA, "Avhrr." `https://www.earthdata.nasa.gov/sensors/avhrr`, 2000. Accessed: 2023-6-13.

[89] planet, "planetscope." `https://www.planet.com/?utm_source=google&utm_medium=paid-search&utm_campaign=discovery&utm_content=pros-leads-responsive-search-0623&utm_`

source=google&utm_medium=paid-search&gad=1, 2019. Accessed: 2023-5-18.

[90] ISRO, "cartosat." https://www.isro.gov.in/CARTOSAT_1.html, 2019. Accessed: 2023-5-18.

[91] J. Fan, J. Bai, Z. Li, A. Ortiz-Bobea, and C. P. Gomes, "A gnn-rnn approach for harnessing geospatial and temporal information: application to crop yield prediction," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, pp. 11873–11881, 2022.

[92] A. de Wit, G. Duveiller, and P. Defourny, "Estimating regional winter wheat yield with wofost through the assimilation of green area index retrieved from modis observations," *Agricultural and forest meteorology*, vol. 164, pp. 39–52, 2012.

[93] A. Kaur, P. Goyal, R. Rajhans, L. Agarwal, and N. Goyal, "Fusion of multivariate time series meteorological and static soil data for multistage crop yield prediction using multi-head self attention network," *Expert Systems with Applications*, vol. 226, p. 120098, 2023.

[94] E. He, Y. Xie, L. Liu, W. Chen, Z. Jin, and X. Jia, "Physics guided neural networks for time-aware fairness: An application in crop yield prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 14223–14231, 2023.

[95] T. Jiang, M. Huang, I. Segovia-Dominguez, N. Newlands, and Y. R. Gel, "Learning space-time crop yield patterns with zigzag persistence-based lstm: Toward more reliable digital agriculture insurance," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 12538–12544, 2022.

[96] X. Xiao, T. He, S. Liang, X. Liu, Y. Ma, S. Liang, and X. Chen, "Estimating fractional snow cover in vegetated environments using modis surface reflectance data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 114, p. 103030, 2022.

[97] I. Jebli, F.-Z. Belouadha, M. I. Kabbaj, and A. Tilioua, "Prediction of solar energy guided by pearson correlation using machine learning," *Energy*, vol. 224, p. 120109, 2021.

[98] Ü. Ağbulut, A. E. Gürel, and Y. Biçen, "Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison," *Renewable and Sustainable Energy Reviews*, vol. 135, p. 110114, 2021.

[99] A. Sharafati, K. Khosravi, P. Khosravinia, K. Ahmed, S. A. Salman, Z. M. Yaseen, and S. Shahid, "The potential of novel data mining models for global solar radiation prediction," *International Journal of Environmental Science and Technology*, vol. 16, pp. 7147–7164, 2019.

[100] S. Belaid and A. Mellit, "Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate," *Energy Conversion and Management*, vol. 118, pp. 105–118, 2016.

[101] D. Díaz-Vico, A. Torres-Barrán, A. Omari, and J. R. Dorronsoro, "Deep neural networks for wind and solar energy prediction," *Neural Processing Letters*, vol. 46, pp. 829–844, 2017.

[102] J. M. Barrera, A. Reina, A. Maté, and J. C. Trujillo, "Solar energy prediction model based on artificial neural networks and open data," *Sustainability*, vol. 12, no. 17, p. 6915, 2020.

[103] F. Rodríguez, A. Fleetwood, A. Galarza, and L. Fontán, "Predicting solar energy generation through artificial neural networks using weather forecasts for microgrid control," *Renewable energy*, vol. 126, pp. 855–864, 2018.

[104] modis, "Modis product." `https://developers.google.com/earth-engine/datasets/catalog/MODIS_006_MOD10A1`. [Accessed 06-January-2023].

[105] V. C. Corporation, "Weather Data Services — Visual Crossing — visualcrossing.com." `https://www.visualcrossing.com/weather/weather-data-services#`. [Accessed 04-june-2022].

[106] K. Gavahi, P. Abbaszadeh, and H. Moradkhani, "Deepyield: A combined convolutional neural network with long short-term memory for crop yield forecasting," *Expert Systems with Applications*, vol. 184, p. 115511, 2021.

[107] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 92–93, 2020.

[108] S. Skakun, B. Franch, E. Vermote, J.-C. Roger, C. Justice, J. Masek, and E. Murphy, "Winter wheat yield assessment using landsat 8 and sentinel-2 data," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5964–5967, IEEE, 2018.

[109] J. Wu, Q. Cheng, H. Li, S. Li, X. Guan, and H. Shen, "Spatiotemporal fusion with only two remote sensing images as input," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 6206–6219, 2020.

[110] N. J. Pastick, B. K. Wylie, and Z. Wu, "Spatiotemporal analysis of landsat-8 and sentinel-2 data to support monitoring of dryland ecosystems," *Remote sensing*, vol. 10, no. 5, p. 791, 2018.

[111] Q. Wang, G. A. Blackburn, A. O. Onojeghuo, J. Dash, L. Zhou, Y. Zhang, and P. M. Atkinson, "Fusion of landsat 8 oli and sentinel-2 msi data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3885–3899, 2017.

[112] Z. Shao, J. Cai, P. Fu, L. Hu, and T. Liu, "Deep learning-based fusion of landsat-8 and sentinel-2 images for a harmonized surface reflectance product," *Remote Sensing of Environment*, vol. 235, p. 111425, 2019.

[113] A. Orynbaikyzy, U. Gessner, B. Mack, and C. Conrad, "Crop type classification using fusion of sentinel-1 and sentinel-2 data: Assessing the impact of feature selection, optical data availability, and parcel sizes on the accuracies," *Remote sensing*, vol. 12, no. 17, p. 2779, 2020.

[114] B. Chen, J. Li, and Y. Jin, "Deep learning for feature-level data fusion: Higher resolution reconstruction of historical landsat archive," *Remote Sensing*, vol. 13, no. 2, p. 167, 2021.

[115] M. M. Rahman and A. Robson, "Integrating landsat-8 and sentinel-2 time series data for yield prediction of sugarcane crops at the block level," *Remote Sensing*, vol. 12, no. 8, p. 1313, 2020.

[116] T. Dong, J. Liu, B. Qian, L. He, J. Liu, R. Wang, Q. Jing, C. Champagne, H. McNairn, J. Powers, *et al.*, "Estimating crop biomass using leaf area index derived from landsat 8 and sentinel-2 data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 168, pp. 236–250, 2020.

[117] B. Ping, Y. Meng, and F. Su, "An enhanced linear spatio-temporal fusion method for blending landsat and modis data to synthesize landsat-like imagery," *Remote Sensing*, vol. 10, no. 6, p. 881, 2018.

[118] Z. Tan, M. Gao, J. Yuan, L. Jiang, and H. Duan, "A robust model for modis and landsat image fusion considering input noise," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[119] Q. Wang, Y. Zhang, A. O. Onojeghuo, X. Zhu, and P. M. Atkinson, "Enhancing spatio-temporal fusion of modis and landsat data by incorporating 250 m modis data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 9, pp. 4116–4123, 2017.

[120] J. Wei, L. Wang, P. Liu, X. Chen, W. Li, and A. Y. Zomaya, "Spatiotemporal fusion of modis and landsat-7 reflectance images via compressed sensing," *IEEE*

Transactions on Geoscience and Remote Sensing, vol. 55, no. 12, pp. 7126–7139, 2017.

[121] S. Bouabid, M. Chernetskiy, M. Rischard, and J. Gamper, "Predicting landsat reflectance with deep generative fusion," *arXiv preprint arXiv:2011.04762*, 2020.

[122] J. Schreier, G. Ghazaryan, and O. Dubovyk, "Crop-specific phenomapping by fusing landsat and sentinel data with modis time series," *European Journal of Remote Sensing*, vol. 54, pp. 47–58, 2021.

[123] W. C. Sleeman IV, R. Kapoor, and P. Ghosh, "Multimodal classification: Current landscape, taxonomy and future directions," *ACM Computing Surveys*, vol. 55, no. 7, pp. 1–31, 2022.

[124] X. Chen, N. Zhang, L. Li, S. Deng, C. Tan, C. Xu, F. Huang, L. Si, and H. Chen, "Hybrid transformer with multi-level fusion for multimodal knowledge graph completion," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 904–915, 2022.

[125] D. Fan, L. Wan, W. Xu, and S. Wang, "A bi-directional attention guided cross-modal network for music based dance generation," *Computers and Electrical Engineering*, vol. 103, p. 108310, 2022.

[126] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[127] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of IEEE/CVF international conference on computer vision*, pp. 357–366, 2021.

[128] A. Shah, S. Sra, R. Chellappa, and A. Cherian, "Max-margin contrastive learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 8220–8230, 2022.

[129] A. T. Tefera, B. P. Banerjee, B. R. Pandey, L. James, R. R. Puri, O. Cooray, J. Marsh, M. Richards, S. Kant, G. J. Fitzgerald, *et al.*, "Estimating early season growth and biomass of field pea for selection of divergent ideotypes using proximal sensing," *Field Crops Research*, vol. 277, p. 108407, 2022.

[130] M. D. Johnson, W. W. Hsieh, A. J. Cannon, A. Davidson, and F. Bédard, "Crop yield forecasting on the canadian prairies by remotely sensed vegetation indices and machine learning methods," *Agricultural and forest meteorology*, vol. 218, pp. 74–84, 2016.

[131] A. Nagy, A. Szabó, O. D. Adeniyi, and J. Tamás, "Wheat yield forecasting for the tisza river catchment using landsat 8 ndvi and savi time series and reported crop statistics," *Agronomy*, vol. 11, no. 4, p. 652, 2021.

[132] A. Vannoppen and A. Gobin, "Estimating farm wheat yields from ndvi and meteorological data," *Agronomy*, vol. 11, no. 5, p. 946, 2021.

[133] S. A. Shammi and Q. Meng, "Use time series ndvi and evi to develop dynamic crop growth metrics for yield modeling," *Ecological Indicators*, vol. 121, p. 107124, 2021.

[134] M. U. Liaqat, M. J. M. Cheema, W. Huang, T. Mahmood, M. Zaman, and M. M. Khan, "Evaluation of modis and landsat multiband vegetation indices used for wheat yield estimation in irrigated indus basin," *Computers and Electronics in Agriculture*, vol. 138, pp. 39–47, 2017.

[135] J.-P. Dedieu, B. Z. Carlson, S. Bigot, P. Sirguey, V. Vionnet, and P. Choler, "On the importance of high-resolution time series of optical imagery for quantifying the effects of snow cover duration on alpine plant habitat," *Remote Sensing*, vol. 8, no. 6, p. 481, 2016.

[136] S. Gascoin, M. Grizonnet, M. Bouchet, G. Salgues, and O. Hagolle, "Theia

snow collection: high-resolution operational snow cover maps from sentinel-2 and landsat-8 data, earth syst. sci. data, 11, 493–514," 2019.

[137] S. Xiang, M. Wang, J. Xiao, G. Xie, Z. Zhang, and P. Tang, "Cloud coverage estimation network for remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, 2023.

[138] A. Baran, S. Lerch, M. El Ayari, and S. Baran, "Machine learning for total cloud cover prediction," *Neural Computing and Applications*, vol. 33, no. 7, pp. 2605–2620, 2021.

[139] G. Andrianakos, D. Tsourounis, S. Oikonomou, D. Kastaniotis, G. Economou, and A. Kazantzidis, "Sky image forecasting with generative adversarial networks for cloud coverage prediction," in *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1–7, IEEE, 2019.

[140] B.-Y. Kim, J. W. Cha, and K.-H. Chang, "Twenty-four-hour cloud cover calculation using a ground-based imager with machine learning," *Atmospheric Measurement Techniques*, vol. 14, no. 10, pp. 6695–6710, 2021.

[141] Y. Son, Y. Yoon, J. Cho, and S. Choi, "Cloud cover forecast based on correlation analysis on satellite images for short-term photovoltaic power forecasting," *Sustainability*, vol. 14, no. 8, p. 4427, 2022.

[142] Y. Cai, W. Zheng, X. Zhang, L. Zhangzhong, and X. Xue, "Research on soil moisture prediction model based on deep learning," *PloS one*, vol. 14, no. 4, p. e0214508, 2019.

[143] Q. Li, Z. Li, W. Shangguan, X. Wang, L. Li, and F. Yu, "Improving soil moisture prediction using a novel encoder-decoder model with residual learning," *Computers and Electronics in Agriculture*, vol. 195, p. 106816, 2022.

[144] Q. Li, Y. Zhu, W. Shangguan, X. Wang, L. Li, and F. Yu, "An attention-aware lstm model for soil moisture and soil temperature prediction," *Geoderma*, vol. 409, p. 115651, 2022.

[145] E. Taktikou, G. Bourazanis, G. Papaioannou, and P. Kerkides, "Prediction of soil moisture from remote sensing data," *Procedia engineering*, vol. 162, pp. 309–316, 2016.

[146] J. Zhao, Y. Dong, M. Zhang, and L. Huang, "Comparison of identifying land cover tempo-spatial changes using globcover and mcd12q1 global land cover products," *Arabian Journal of Geosciences*, vol. 13, pp. 1–12, 2020.

[147] NASA-USDA, "Smap." `https://developers.google.com/earth-engine/datasets/catalog/NASA_USDA_HSL_SMAP10KM_soil_moisture`. [Accessed 07-November-2023].

[148] modis, "Modis product." `https://developers.google.com/earth-engine/datasets/catalog/MODIS_061_MCD12Q1`. [Accessed 26-November-2023].

[149] Z. Gao, Y. Wu, X. Zhang, J. Dai, Y. Jia, and M. Harandi, "Revisiting bilinear pooling: A coding perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 3954–3961, 2020.

[150] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu, "Disentangled representation learning," *arXiv preprint arXiv:2211.11695*, 2022.

[151] E. Rolf, K. Klemmer, C. Robinson, and H. Kerner, "Mission critical–satellite data is a distinct modality in machine learning," *arXiv preprint arXiv:2402.01444*, 2024.

[152] C. Nast, "Datasets for deep learning — wired.com." `https://www.wired.com/beyond-the-beyond/2020/02/datasets-deep-learning/`. [Accessed 22-03-2024].

[153] L. Meng, H. Li, B.-C. Chen, S. Lan, Z. Wu, Y.-G. Jiang, and S.-N. Lim, "Adavit: Adaptive vision transformers for efficient image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12309–12318, 2022.

[154] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, "Dynamicvit: Efficient vision transformers with dynamic token sparsification," *Advances in neural information processing systems*, vol. 34, pp. 13937–13949, 2021.

[155] L. Hou, R. Y. Pang, T. Zhou, Y. Wu, X. Song, X. Song, and D. Zhou, "Token dropping for efficient bert pretraining," *arXiv preprint arXiv:2203.13240*, 2022.

[156] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10078–10093, 2022.

[157] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao, "Videomae v2: Scaling video masked autoencoders with dual masking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14549–14560, 2023.

[158] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

List of research publications

Conference/Journal Papers:

1. Poonam Goyal, Arshveer Kaur, Arvind Ram, and Navneet Goyal, "Efficient Representation Learning of Satellite Image Time Series and Their Fusion for Spatiotemporal Applications", in 38th AAAI Conference on Artificial Intelligence (AAAI 2024).

2. Arshveer Kaur, Poonam Goyal, and Navneet Goyal, "LSFuseNet: Dual-Fusion of Landsat-8 and Sentinel-2 Multispectral Time Series for Permutation Invariant Applications", in IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA 2023).

3. Arshveer Kaur, Poonam Goyal, Rohit Rajhans, Lakshya Agarwal, and Navneet Goyal, "Fusion of multivariate time series meteorological and static soil data for multistage crop yield prediction using multi-head self attention network", in Expert Systems with Applications (2023). (Impact Factor: 8.2)

4. Arshveer Kaur, Poonam Goyal, Kartik Sharma, Lakshay Sharma, and Navneet Goyal, "A generalized multimodal deep learning model for early crop yield prediction", in IEEE International Conference on Big Data 2022.

Communicated Papers:

1. Arshveer Kaur, Poonam Goyal, Vansh Bansal, Deep Pandya, and Navneet Goyal, "SpInN: A Pre-trained Model for Spectral Indices Recommendation for Earth Observation Applications using Satellite Data", in IEEE Transactions on Neural Network and Learning System. [Communicated]

2. Arshveer Kaur, Poonam Goyal, Niranjan, and Navneet Goyal, ”SatTran: A transformer for satellite image time series exploiting spatiotemporal redundancies”, in NeurIPS 2024 [Communicated]

3. Arshveer Kaur, Poonam Goyal, and Navneet Goyal, ”A review on prediction through image time series for permutation invariant applications” [to be Communicated shortly]

Patents:

1. Provisional Patent Filed ”A System And Method For Processing High-Resolution Satellite Image Time Series (SITS) Through A PatchNet Model”
   Application no. 202411011144 dated 17 February 2024.

2. Patent Filing in process for ”SaTran: A transformer for satellite image time series”.

Biography of the Candidate

Arshveer Kaur is currently a full-time research scholar in the Department of Computer Science & Information Systems at Birla Institute of Technology & Science, Pilani, Rajasthan, India. During Ph.D., she has been working as a teaching assistant for courses like Computer Programming, Data Mining, Foundation of Data Science, etc. She received a Scholarship from ACM-W and a travel grant from AAAI for presenting her paper in 38th AAAI conference on Artificial Intelligence (AAAI 2024) held in Vancouver, Canada. She has 10 publications in total. She has completed her M. Tech. (Computer Engineering) in 2016 from Punjab Engineering College (PEC) Chandigarh and B. Tech. (Computer Science & Engineering) in 2014 from Punjab Technical University, Punjab. She worked in TCS as an Assistant Software Engineer before joining Ph.D. Her research interests are Data Mining, Machine Learning, Data Analytics, etc.

Contact her at p2017432@pilani.bits-pilani.ac.in

Biography of the Supervisor

Prof. Navneet Goyal is currently a Senior Professor and Head of the Department in the Department of Computer Science and Information Systems at Birla Institute of Technology and Science Pilani, Pilani campus. He did his PhD from IIT Roorkee in 1995 and joined BITS-Pilani in the same year. He has 29 years of research and teaching experience at BITS Pilani.

His main research area is Artificial Intelligence & Machine Learning. His current research focus is on Generative AI and AI for Earth Observation. His recent research focuses on high resolution satellite image time series analytics for various earth observation applications like crop yield prediction, snow cover prediction, solar energy prediction, etc. He is an active member of APPCAIR – an alumni-funded AI research center. He is the coordinator of Disruptive Technologies Lab. which focuses on problems at the confluence of AI, IoT, and Blockchain technologies. He has also done extensive work in Big Data Analytics & High-Performance Computing.

He is a recipient of two globally competitive research awards from IBM & Google. In 2010 he received IBM's global Scalable Data Analytics Research Innovation Award under their "Smarter Planet" initiative. The award is in the area of Agriculture. He also received the prestigious Google Impact Scholar award under their "AI for Social Good" initiative in 2021. The award is for a project about reducing wildlife crime using AI.

In March 2024, he received the first phase funding of INR 2 crores from the Ministry of Education, GoI for setting up a CoE for AI in Agriculture. He is the Chief Program Manager of a consortium of nine organizations led by BITS Pilani which is working on a project on use of AI and Satellite Technology for ensuring food security. He recently got a project fund of INR 1.50 crores from the Ministry of Tribal Affairs, GoI in which he and his team will work on AI-intervened tribal development.

He is also leading multiple research collaborations with the industry in the area of

Generative AI, the two most prominent being with Mercedes-Benz Micron Technologies. More on his research contributions can be found at

`https://www.bits-pilani.ac.in/pilani/navneet-goyal/`

Contact him at `goel@pilani.bits-pilani.ac.in`.