

Chapter 2

Literature Review

2.1 Introduction

The United Nations 17 Sustainable Development Goals (SDGs) and 230 indicators listed in the 2030 Agenda for sustainable development objectifies establishing a neat balance between development and the importance of ensuring wellness of the environment [36]. Major inclusion in SDG emphasizes on urban settlements to be more inclusive, safe for inhabitants, resilient to various influences, sustainable and make sustainable use of resources in the ecosystem, facilitate restoration of greenary, manage forest cover, prevent desertification of greenspaces, rejuvenate degradation of land, and prevent biodiversity loss. [25].

Developmental initiatives should be progressive, sustainable, and prosperous. It should preserve the very essence of ecosystem services. Ecosystem services are essential for the welfare and wellbeing of human kind as well as nature as it provides all essential ingredients that shape and save the both. It provides services related to provisions, regulations, culture and support [26]. Further, ecosystem disservices would result in degradation of ecosystem and subsequently hamper ecosystem services [27] resulting in disturbance and distress in the ecosystem. Achievement of sustainable development goals mandates efficient maintenance of the sustainable development goals.

The United Nations' 2014 report, "*World Urbanization Prospects*" states that urbanization has been accelerating in recent decades, frequently doubling in size every ten years. By 2050, there could be an additional 2.5 billion urban residents, with nearly 90% of this growth taking place in Asia and Africa. Developing nations like China and India will experience the fastest urbanization. In addition to posing problems for infrastructure, housing, and employment, this extraordinary urban population growth may also put further strain on urban land management, spatial equity, and more generally sustainable development [28].

In recent years, urbanization has intensified in emerging nations, and fair and sustainable urban development has gained significance in various states and countries worldwide. As global urbanization escalates, the focus on equity and sustainability in urban environments is intensifying among researchers, academics, and practitioners involved in urban planning, sustainable development and public policy across various countries, including the United States, China, India, and the United Kingdom.

Studies related to land use transformations and transition have been to a great extent made possible with significant technological advancement in the field of geospatial sciences such as GIS, RS and Global Positioning System (GPS) enabling researcher to efficiently and effectively monitor and study land use and plan assertive measure for assisting responsible agencies to make real time quality decisions.

Remotely sensed data can be extensively analyzed to estimate population [37], understand the nature of sprawl, climate change [38] deforestation [39], urban sprawl [40], cultivation patterns [41], land use transformation and economy. GIS facilitates

real-time data analysis, efficient decision making, effective data visualization and representation and optimizes results reducing cost. GPS facilitates efficient and accurate localization, ensures global coverage and is highly effective.

Research initiatives motivated towards study of geomorphological trends and patterns neatly integrates these technological solutions to facilitate fast and efficient analysis with greater precision of measurement and representation to assist planner and authority to pursue quality decision for implementing a sustainable plan.

The effectiveness of the results obtained through use of geospatial solutions can be further magnified, enhanced, verified and validated by combining it with the finding of ground truthing initiatives such as socioeconomic surveys, censuses, and any other form of biophysical information acquisition tool deployed for understanding the land use trends and factors influencing the same [41] Feeds from social media applications may be further combined to build an understanding of the nature of human activities [38] and its impact on land use [43].

Remotely sensed satellite image may be further combined with data from social media forum for identifying urban sprawl and determine its influence on sustainable urban growth and ecosystem service. Such effort may be realizable only in situation if the social media data are available for use.

Liu et. al. [44] assessed worldwide urban sprawl at the metropolitan level utilizing remote sensing data and census statistics from 2000 and 2010. The findings demonstrate that, on a global scale, Metropolitan Areas (MAs) exhibited increased

sprawl characteristics in 2010 compared to 2000. As investment increased, urban sprawl intensified. Furthermore, from a dynamic standpoint, all metropolitan areas expanded extensively in comparison to the patterns observed in the base year of 2000. The findings elucidate the mechanisms of urban development and concurrently aid urban planners in evaluating the possible hazards of urban sprawl and ineffective development throughout various phases of urban growth. Yet, the analysis confirms that, when compared to the primary measurement of density, the multidimensional measure of sprawl offers a number of benefits for characterizing sprawl, a conclusion that is corroborated by Downs [45]. Multidimensional measurements of urban sprawl are useful because it takes into account both dynamic and static patterns, as opposed to just low-density urban patterns. Nonetheless, this worldwide index possesses a deficiency.

Espindola et. al. examined the urban spread of Teresina, Brazil over the past forty years (1974-2014) along with the social and environmental ramifications of these patterns [46]. The findings demonstrated that urban sprawl has been persistent over the years, and significantly accelerated during the interval 2010-2014. Areas with high demographic densities were typically found in outlying locations where the majority of people were in low-income category. The population density differs throughout the city and significantly deviates from the standards established by urban legislation. The land use maps were derived from Landsat images taken according to seasonal parameters between June and August, with cloud cover not exceeding 10%. The utilized images are Thematic Mapper (TM) for the years 1990, 2000, and 2010, and Operational Land Imager (OLI) for 2014. The land categorization utilized the supervised Bhattacharya algorithm executed in SPRING program [47]. This method

classifies the imagery based on segmented images, indicating that the objects for classification are areas rather than individual pixels of the image. The region-growing technique for picture segmentation, included in SPRING program, was employed. A segmented image, was used to prevent fragmented classes with excessive pixel variation inside a confined area [48].

This research initiative [49] examined the correlation between development and alterations in land use, as well as their effects on the landscape of Fez, Morocco highlighting the significance of the expansion of impervious surfaces. Satellite imagery and data from the census was utilized to discern various patterns of land use change and urban expansion from 1984 to 2013. The classification and assessment of satellite imagery were conducted with Erdas Imagine and ArcGIS software. The evaluation of urban sprawl in Fez occurred during a span of 29 years (1984–2013). The total precision of land use change maps, produced by post-classification change detection methods and assessed via multiple ways, varied between 78% and 87%.

Effat et. al. used [29], Multi-Temporal Landsat TM and ETM imagery obtained between 1984 and 2013 in conjunction with Shannon's entropy to try to comprehend, identify, and measure the spatial arrangement of Cairo's, Egypt urban growth. Supervised classification was utilized to delineate built-up regions and assess alterations in urban land-use categories across the municipal wards. Shannon's entropy was utilized for modeling the urban sprawl, trends, and spatial transformations of the metropolis. Entropy data for the city's electoral constituencies were modeled and utilized in an interpolation function to generate an entropy surface (index) for each obtained temporal image. This index illustrates the spatial distribution of urban

expansion and offers a visual comparison of the entropy phenomena in these wards. Results suggest that Shannon's entropy index rose from 1.4615 in 1984 to 2.1023 in 2013, signifying more scattered urban growth, indicative of urban sprawl. A regression study was conducted on the population growth rate and the developed areas. The findings elucidate the sprawl patterns and dynamics within Cairo's electoral wards, offering a visual comparison.

2.2 Urban sprawl, its types, characteristics, causes, consequences and remedial measures

The research pertaining to the development of sustainable communities is termed Sprawl [8]. Urban sprawl is characterized by low-density, chaotic, uncontrolled, unplanned, and unrestricted expansion of urban areas into rural and natural landscapes to accommodate the rising demand driven by population growth and aspirations for economic stability and prosperity.

The community's effort towards socio-economic advancement does invite changes in environmental conditions. The research concerning the development of sustainable communities is termed Sprawl [8]. Urban sprawl is characterized by low-density, chaotic, uncontrolled, unplanned, and unrestricted expansion of urban areas into rural and natural landscapes, driven by the rising demand stemming from population growth and the pursuit of economic stability and prosperity.

The ongoing population growth in urban areas necessitates an expansion of infrastructure to address the emerging demands associated with education, governance, transportation, recreation, public health, and convenience. Urban sprawl

leads to adverse effects, including the degradation of agricultural land, open spaces, and ecologically sensitive areas [16]. Moreover, sprawl pertains to the extension of a urban cores radially or linearly. In essence, when the population of a neighborhood or city increases, the town's boundaries extend to accommodate this expansion. Sprawls generally manifest on the peripheries of urban regions, either at the city's boundary or adjacent to principal thoroughfares.

2.2.1 Types of urban sprawl

Sprawl may be of different types [1], some of which are leapfrog sprawl, ribbon sprawl (also known as strip sprawl), low-density single-use sprawl, exurban, and edge cities.

- **Leapfrog sprawl** has clusters of infrastructural setups sparsely scattered over open spaces. Such organizations may invite increased cost investments based on the need to establish and facilitate communication.
- **Ribbon sprawl** (also known as strip sprawl) is where development happens along linear features such as transportation and river networks. Such organization may invite problems related to social order (including safety and security), traffic congestion, and environmental hazards.
- **Low-density single-use sprawl** is a type of sprawl where dedicated zones are created for various purposes such as industries, residences, health, administration, and education. Although effective, such organization invites increased cost for development as well as calls for increased reliance on transportation aids.
- **Exurban sprawl** is usually observed along the peripheries of an urban core encroaching upon the rural spaces. Some of the prime reasons for developments may be the reduced cost and availability of ample spaces free of problems such as overcrowding and different forms of pollution.

- **Edge cities** are suburban, self-reliant, and self-contained settlements far from the urban core established as an alternative to saturated urban cores. Such developments mandate investment in transportation facilities to connect them to the urban cores.

It can be inferred that process of urbanization occurs either in a radial pattern around a well-established metropolis or linearly along highways during a specific timeframe [10]. Sprawling areas exhibit two distinct map patterns: radial and linear. Sudhira et al. [10] assert that to comprehend the intricacy of urban sprawl, it is imperative to conduct analyses on land use change and recognize patterns of urban growth. Within the realm of literature, there is a lack of clarity the distinction between urban growth and urban sprawl, but it is implied that urban sprawl might be considered a form of urban growth.

2.2.2 Characteristics of urban sprawl

Urban sprawl possesses multifaceted characteristics [18]. The following characteristics are delineated:

- **Density:** Denotes the quantity of dwelling units per area of unit. The construction of single-family residences, or housing units designed for fewer occupants, would significantly exacerbate urban sprawl, resulting in the consumption of substantial areas of previously productive or green land [19].
- **Continuity:** Denotes growth patterns that are erratic, discontinuous, dispersed, scattered, and sparsely situated (not in proximity to existing urban cores), resulting in more expenditure of effort, money, and land. Developers refer to this form of development as leapfrog development [20] [21].

- **Concentration:** Indicates low density, reduced compactness, spiraling, and dispersed characteristics demonstrated by developments extending from the urban core [50]. This type of development requires increased labor, resources, and land. Additionally, this type of development requires investments in constructing connectivity infrastructure.
- **Clustering:** Denotes enclaves of urbanization characterized by disjointed urban configurations stemming from inefficient and disorganized residential and commercial infrastructure, leading to suboptimal utilization of land, infrastructure, and transportation. In this context, developments are fragmented and unevenly allocated, disregarding closeness and access to key services [19].
- **Centrality:** Indicates the efficacy of urban development in relation to a predominant urban center. Weak centrality is associated with inadequate connectivity to urban centers, resulting from the dispersal of services and job opportunities, which fosters multifaceted growth away from the metropolitan core. This would undoubtedly require additional land, necessitate the formation of further infrastructure, the construction of transportation systems, an increased dependence on automotive transit, and result in inefficient public transportation.
- **Nuclearity:** Indicates the existence of a clearly defined and systematically organized metropolitan core. In urban areas, low nuclearity denotes the lack of a dominant urban core, resulting in low-density, poorly organized, and inadequately connected polycentric suburbs with inefficient accessibility.
- **Proximity:** Indicates the nearness of diverse land use patterns, delineated by stringent zoning regulations, including areas designated for labor, commerce, education, and healthcare. Poor proximity in sprawl patterns will necessitate

additional effort for the creation of transportation facilities and lead to increased reliance on automobile transport [18].

- **Mixed Uses:** Denotes the amalgamation of several land use categories, including institutional, residential properties, and commercial developments. Mixed-use sprawl will result in traffic congestion, elevated infrastructural expenses, and social security challenges.
- **Strip Development:** Denotes the radial spread of infrastructure emanating from an urbanized core [22]. Strip expansion is frequently driven by economic advantages and commercial interests.
- **Indistinct boundaries among urban and rural regions:** Illustrates sprawl patterns that intrude upon or advance toward rural peripheries due to developments emanating from metropolitan centers [23].

2.2.3 Causes of urban growth and sprawl

Multiple factors, including economic, social, cultural, ecological, and institutional policies and frameworks, often influence urban sprawl.

Below are a few significant causes:

- **Opportunities for socio-economic advancement:** The presence of prospects for socio-economic advancement and growth is the primary catalyst that drives urban migration, hence necessitating urban expansion.
- **Pursuit of enhanced quality of life:** A prevalent belief among the populace is that suburban living provides superior comfort, characterized by less pollution and the availability of green spaces, hence encouraging outward expansion.
- **Growing population:** The potential for societal and economic advancement frequently drives individuals to relocate to urban areas. Migration can occur both

interstate and across state lines, contingent upon the availability of opportunities. This extraordinary temporal demand necessitates urban expansion due to the distress caused by urban overcrowding. Consequently, people view formerly undeveloped periphery areas as a viable option for expansion to meet the accommodation needs of the increasing population.

- **Institutional requirements stemming from socio-economic affluence:** Economic prosperity enhances affordability, necessitating the creation of facilities for public convenience, including parks, sports complexes, retail spaces, educational institutions, administrative facilities, and residential areas. Urban areas may not address these rising significant needs that drive urban expansion.
- **Escalating desire for affordable housing:** Housing expenses in urban areas are beyond the financial capacity of the majority of urban residents. Consequently, there is a heightened demand for inexpensive housing, creating potential for development in areas situated radially or linearly around metropolitan centers, which leads to urban expansion [24].
- **The abundance of cheap land for expansion in suburban areas:** Given that land prices in urban areas may be prohibitively high, the accessibility of land in radial or linear proximity to these centers presents a cost-effective alternative for triggering development projects, which is a primary factor contributing to urban [51]. The public typically prefers pleasant residences in suburban areas over crowded metropolitan environments [24].
- **Availability of transportation:** The availability of transportation infrastructure in the suburbs enhances access to inexpensive commuting options from the outskirts to the urban core, thereby facilitating urban expansion and promoting developmental projects in peripheral areas. [26][27].

- **Infrastructural initiatives:** Infrastructural initiatives promote commercial, residential, and institutional prosperity, offering useful insights for enhancements and future opportunities. The availability of economical transportation and operational expenses, such as fuel costs, accelerates expansion.
- **The impact of socio-economic differentiation and lifestyle preferences:** Various socio-economic factors have been identified as significant drivers of in-migration and migration out from urban centers, contributing to urban expansion. These factors include a) income [28], b) racial conflict [29], c) age [30], and d) social clusters.
- **Zoning Regulations and Land Utilization Policies:** Unlike standard laws on zoning (such as Euclidean Zoning) that necessitate clear separation between industrial, residential properties and commercial land uses, the conventional mixed-use approach advocates for the multifunctional use of land. This facilitates a decrease in friction costs, including transportation expenses, while also encouraging sustainable practices.
- **Unsustainable urbanization planning and inadequate institutional coordination:** The unchecked expansion of metropolitan areas into underdeveloped or undeveloped regions frequently stems from insufficient and poorly conceived urban development plan, temporal requirements, which are often reactive, primarily instigate this.ve. This method of urban development is typically unmanageable and frequently imperceptible. Leapfrog Development serves as a prominent illustration.
- **Institutional incentives:** Institutional incentives and actions that facilitate urban expansion include the establishment of transportation, water, and power networks.

Similarly, affordable housing initiatives, reasonable taxation, subsidies, and zoning regulations also facilitate urban expansion.

- **Technological advancements:** Technological innovations have rendered solutions such as the World Wide Web both accessible and affordable. Moreover, due to the decentralization of labor arrangements, metropolitan residents increasingly favor living in fringe cities.
- **Impact of globalization and uniform developmental practices:** Development of urban habitats at the expense of flourishing rural landscapes to address the evolving and significant demands of globalization.

2.2.4 Consequences of urban sprawling

The following delineates the detrimental effects of urban sprawl concerning economic, social, cultural, ecological, and health issues:

i) Social concerns

- **Unregulated population growth:** Urban sprawl attracts an influx due to advancing developmental activities.
- **Social pollution:** The inculturation and cultural erosion stemming from international influences, security issues, and many social concerns.
- **Scarcity of resources:** Urbanization initiatives may encroach upon productive agricultural land, resulting in a depletion of local resources. Similarly, excessive usage of vital resources may lead to shortages.
- **Social stratification:** Community may be divided into income classes, races, and religions, resulting in a distressed and troubled civil society.
- **Deteriorating social relationships:** Increased reliance on technical solutions may erode social ties, hence hindering social integration.

- Access to important services: The lack of educational institutions, healthcare facilities, recreational areas, retail spaces, and transportation options in low-density projects hinders access to essential services.

ii) Environmental concerns

- Escalated land consumption: Property is being utilized at an unparalleled rate to satisfy the demands of a growing population. The widespread implementation of leapfrog patterns of growth would lead to inefficient land use, resulting in ecological issues.
- Environmental degradation: Encompasses the loss of local biodiversity and the disruption or destruction of habitats.
- Noise, air, water, and soil pollution: A growing population, heightened automobile usage, industrialization, drainage issues, and inadequate waste management contribute to noise, environmental, and water pollution.
- Climate change: Rise in temperature attributable to greenhouse gas releases from suboptimal systems.
- The loss of natural resources: Reduction in groundwater reserves due to inadequate surface water percolation, loss of green areas, and other natural sources.
- Deteriorating water quality: Surface runoff from metropolitan areas may transport dangerous pollutants, consequently compromising the quality of subterranean water sources.
- Elevated flood risk: Accelerated urbanization may result in surface sealing, causing swift surface water flow and producing flood conditions.
- Degradation of arable land: Unregulated and haphazard urban sprawl may intrude upon fertile agricultural land, thereby diminishing productivity.

- Reduction of forested land and waterways: Urban expansion may encroach upon forests and wetlands, so disrupting open spaces and potentially disturbing the local environment and biodiversity.
- Escalated reliance on non-renewable natural resources: The rise in vehicle usage and industrialization will expedite the exhaustion of non-renewable natural resources such as fossil fuels and coal.

iii) Institutional concerns

- Impact on infrastructure: Traffic congestion, heightened power and water usage, and improved waste and sewage management systems.
- Unregulated and spontaneous urbanization: To address the burgeoning and incessant demands of a rising population.
- Augmented energy consumption: The establishment of low-density, dispersed, and dispersed settlements would inevitably lead to heightened energy usage across all facets, including domestic activities and transportation.
- Urban heat island: Emissions from industrial activities and surface heating retention can lead to the formation of urban heat islands, while additional infrastructure constructed for mitigation may exacerbate the issue.
- Safety and security concerns: The process of urbanization engenders issues such as trafficking, addiction, and social security, necessitating effective and prompt mitigation to ensure the safety and security of citizens.
- Crucial resource management: Guidelines and structures must be effectively implemented to sustain a robust supply chain that meets the escalating demands of the population.

- Regulating and orchestrating growth: Guidelines and frameworks must be firmly created to ensure healthy and sustainable urbanization development.

iv) Health concerns

- Sedentary lifestyle: A sedentary lifestyle may result in health complications.
- Initiation of health disorders resulting from pollution: Medical conditions stemming from several types of pollution.

v) Economic concerns

- Inflation: Increased demand and limited supplies may result in increased cost and expenditure.
- Increased cost of infrastructure: Cost of residential spaces, business spaces, and spaces for offices may increase beyond affordability.
- Increased cost of essential commodities: The cost of essential supplies such as food items, the cost of fuels, water supply, and electricity supplies may also increase.

All these five critical spheres of consequences influence the overall quality of life [30].

2.2.5 Mitigation strategies

- Advocate for and encourage smart growth initiatives by promoting the preservation and sustainable utilization of natural resources, encouraging compact clustered development patterns, and restricting construction in environmentally vulnerable regions through the establishment of conservation belts.
- Enhance the urban plan framework by enforcing stringent zoning restrictions, developing initiatives to conserve brownfields and greenfields, promoting green

infrastructure, establishing a comprehensive road network, and improving commuting options to reduce reliance on automobiles.

- Revitalize current urban areas by implementing sustainable infill development projects, promoting the adaptive reuse of existing and aging infrastructure, planning high-density affordable housing programs, and enhancing public services such as education, transportation, public health, and security.
- Enhance public transportation by upgrading facilities and encouraging usage, establish an inexpensive and secure transport ecosystem, and construct pedestrian pathways while promoting their utilization.
- Enhance public awareness by conducting programs that elucidate the advantages of sustainable development and foster public engagement in the formulation of urban plans.
- Promote sustainable growth by offering subsidies for eco-friendly initiatives and tax advantages for conservation strategies.

In the effort towards attainment of the research objectives specified at serial 1.6 the research imitative reviewed various research articles for establishing a comprehensive understanding of the problem, defining the problem, formulating objective, analysing existing solutions, identifying data sources, conceiving solution strategy, and deciding on the deliverables.

2.3 Image Classification

2.3.1 Supervised, Unsupervised and Hybrid Classification

2.3.1.1 Supervised Classification

The following papers has been referred, where similar type of work has been done with different type of images and for different area of interest. Strategically designed, progressive, and environmentally sustainable urbanization of the landscape would significantly facilitate the development of improved societies, resulting in social and economic prosperity; however, these processes adversely affect Land Use and Land Cover (LULC) [52][53]. The issue escalates when unanticipated and ad hoc temporal resource demands take precedence, leading to heightened developmental activities [54][55][56] disrupts the consistent development cycle, resulting in irregular changes in land use and land cover (LULC) [57][58][59]. [60] have generated an imbalance among urban areas, vegetation, and forest cover, hence impacting terrestrial ecosystems [61]. Such developments negatively impact both local and global climates, resulting in severe consequences for the hydrological cycle [62][63], natural resources, aquatic ecosystems, and greenhouse gas emissions [64].

These activities stem from direct or indirect human intervention [65][66] influenced by governmental policies, demographic requirements, prevailing cultural trends, environmental demands, and commercial prospects [67][68]. The ramifications of these alterations are frequently deemed irreversible, negatively impacting the environment, flora, ecology, carbon cycling, meteorological [69][70], and urban ecosystems [71][72]. This is entirely adverse and harmful to the development of sustainable urban environments [73]. LULC modifications are presently among the

most extensively studied areas for comprehending their effects on the global environment [74].

A comprehensive grasp of population absorptiveness and spatial management is essential for sustainable urban growth [75][76][77]. Motivated by escalating demands, advantageous socioeconomic factors, the natural environment [78][79], and a continually growing urban populace [80][81][82] project that urbanization will reach 70% by 2050 [83]. Urban sprawl, a prominent global anthropogenic phenomenon [84], appears to swiftly consume open spaces and natural vegetation to facilitate the development of public infrastructure, including real estate, road networks, waterways, and drainage and communication systems. These establishments are inherently impermeable and demonstrate exceptionally high density. Urban sprawl is heterogeneous, and its dynamics are contingent upon a nation's pursuit of economic growth [85]. The evaluation of sprawl necessitates the identification, monitoring, and comprehension of appropriate justifications for alterations in LULC throughout time [86][87]. The aim of assessment measures is to identify change and ascertain its nature, magnitude, and spatial distribution [88]. It would indeed be vital in developing effective policies and judgments essential for managing urban sprawl.

Research indicates that anthropogenic activities have altered over fifty percent of the Earth's terrestrial surface [89]. These operations are significantly driven by the necessity for industrialization, enhanced employment prospects, and improved quality of life, ensuring economic success. This development heightens the necessity for the establishment of essential infrastructures, including water and power supply, waste and sewage management, residential and recreational facilities, educational institutions,

trade and commerce centers, local markets, and public health services, leading to swift and unregulated urbanization [90][91][92]. Such activities lead to disruptions in local ecology, vegetation, biodiversity [93], hydrology [94], morphological terrain degradation [95][96], climate, and the social well-being of residents [97][98]. It negatively impacts the landscape's shape and the related natural processes [99].

The Composite Water Management Index (CWMI) report published in August 2019 by Niti Aayog indicates that numerous Indian cities, including Delhi, Bengaluru, and Chennai, may imminently encounter "Day Zero" due to rapid population growth and unchecked urbanization driven by escalating demands [93]. These operations entail the removal of extensive forest cover, increasing the land's susceptibility to human-induced disasters such as landslides, floods, and erosion, hence disrupting the entire ecosystem and rendering it very vulnerable [100]. Emerging issues include the formation of urban heat islands [101], pollution in the air and water [102], challenges in resource management [103], and complications in public administration [104][105], all of which significantly affect the standards of life [106].

Thorough understanding, needs-based assessment, and accurate depiction of changes in land use and land cover are essential for the prompt provision of effective solutions to social, economic, and environmental issues [107][108]. The emergence of geospatial technologies, including RS, GPS, and GIS, along with the accessibility of satellite imagery [109] has enhanced and streamlined the examination and analysis of geographical data.

RS is a cost-effective, appropriate [110], accurate, effective and time-saving geospatial instrument that allows researchers across diverse fields of geological, geographical in

scope and environmental studies to examine, analyze, characterize, assess [111], and evaluate various morphological evolutions, extents, and changes in LULC [112][113][114] utilizing data acquired from spaceborne platforms [115] in real-time [116]. The extensive array of spectral, spatial, and temporal resolutions [117][118] linked to economical remotely observed digital satellite data [119][120] significantly aids in the effective identification of diverse morphological features pertinent to research endeavors aimed at societal welfare. GIS serve as a computational framework facilitating the collection, storage, management, preparation, visualization, and assessment of data, which are essential for LULC change detection, as indicated by DeMers [121], utilizing remotely sensed data [122]. Remote sensing, in conjunction with geographic information systems, has significantly aided researchers in identifying, analyzing, comprehending, and mapping the dynamics of various natural and anthropogenic phenomena, including growth of cities, landslides, floods, and land use/land cover changes, through both supervised and unsupervised spectral grouping of spatio-temporal data obtained via band characterization.

The multi-temporal and multi-spectral satellite data offer enormous opportunities. Researchers would be better able to pinpoint the underlying dynamics and trends in the spatial patterns of urban sprawl with the use of distant temporal analysis [123][124]. Utilizing RS and GIS would make it easier for government and affiliated organizations to comprehend and track the extent of urban sprawl [125], measure land changes, and investigate its impacts [126]. This would allow for the development of successful plans and the effective provision of services for the good of society, saving a great deal of time and money [127]. To effectively monitor various dynamics in LULC [128][129][130] a steady and dependable supply of high-quality data is

necessary for figuring out accurate, precise, and timely data on the changes caused to the geographic distribution of LULC [131][132] over large surfaces. Understanding and tracking LULC changes can also be accomplished using classified and developed earth observation missions in addition to remotely sensed data LULC mapping [133]. For instance, the European Space Agency's Copernicus program's Urban Atlas effort uses high-resolution satellite images to create LULC maps for European cities in order to comprehend and analyze urban areas [134].

For quantitative analysis of LULC changes (including features like towns, wetlands, vegetative areas, barren terrain, and wetland) essential for environmental and socioeconomic applications, purpose-specific image classification methods can be used with multispectral satellite images [135]. Numerous factors contribute to the complexity of classification. According to Ghosh and Siddique [136], these methods can be object-based, supervised, unsupervised, or deep learning-based [137]. Either of any of these crucial features—spectral, radiometric, spatial, or temporal—can serve as the foundation for the outcomes. Selecting a classification model, choosing a training set, preprocessing to reduce sensor-related errors, feature extraction, implementing an approach to classification, post-processing, and, lastly, an accuracy evaluation is all included in the image classification activity [138][139]. A training set representative of all classes must be chosen, a suitable classification model must be chosen, appropriate preprocessing operations must be applied (e.g., image registration, radiometric measurement, atmospheric rectification, and topographic correction) must be applied, variables must be chosen (e.g., multitemporal images, multisensory pictures, and ancillary data must be chosen), a suitable classification technique must be chosen (based on purpose), an acceptable postprocessing tool must be chosen to

improve the excellent results derived from the classification method, and ultimately the results should show a high degree of accuracy.

There are many different kinds of picture categorization techniques, from conventional methods to cutting-edge deep learning techniques. The kind of sensor data used determines how well the chosen classification model and related methods perform [140]. The goal and scale of representation determine which data are chosen for classification; IKONOS and SPOT 5 HRG are used at the local scale, Landsat TM/ETM+ and Terra ASTER are utilized at the regional level, and AVHRR, MODIS, and SPOT are utilized at the global level [138].

Clearly, picture categorization methods fall into one of the two general groups listed below: Based on features, Deep Learning, Transfer Learning, Group, Generative, Graph-Based, Hybrid, and classical Machine Learning are examples of classical approaches [141].

Identifiably, image classification techniques can be broadly categorized into either of the following categories: traditional includes Traditional Machine Learning [142], Feature-Based, Deep Learning, Transfer Learning, Ensemble, Generative, Graph-Based, or Hybrid methods. K-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Decision Trees, Random Forests, Naive Bayes, Principal Component Analysis (PCA), Independent Component Analysis (ICA), Local Binary Patterns (LBP), and Maximum Likelihood are some of the examples of Traditional Machine Learning methods [143]. The feature-based method includes Histogram of Oriented Gradients (HOG) [144], Scale-Invariant Feature Transform (SIFT), and Speeded-Up Robust

Features (SURF). Deep Learning method [145] includes Convolutional Neural Networks (CNNs), it makes use of fine-tuned Deep Learning method for classification of complex datasets [146]. Transfer Learning Bagging and Boosting are types of Ensemble models [135]. Generative models include Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) whereas Graph-Based includes Graph Convolutional Networks (GCNs). Hybrid aims at combining various models based on suitability and requirements.

The selection of an image classification approach depends on attributional criteria. Dependency on training samples determines whether to use Supervised or Unsupervised classification approaches. Dependency on parameters such as the mean vector and covariance matrix determines whether to use Parametric or Non-parametric approaches. Based on the use of pixel information, either Per-pixel, Sub-pixel Object-oriented, or Pre-field approaches may be used.

Use of Hard classifier or Soft approaches may be decided based on the requirement of a definitive decision about landcover. Based on the use of spatial information, Spectral, Contextual, Spectral-Contextual approaches may be decided. Supervised classification approaches include traditional classifiers such as maximum likelihood and minimum distance and advanced classifiers such as artificial neural networks and decision tree classifiers, whereas unsupervised classification approaches include classifiers such as K-means. The parametric approach includes classifiers such as maximum likelihood and linear discriminant analysis, whereas non-parametric approach includes classifiers such as artificial neural networks, support vector machines, and decision tree classifiers. Per-pixel approach includes classifiers such as maximum likelihood,

decision tree, and support vector machines, whereas Sub-pixel approach includes Fuzzy-set classifiers. Hard approach includes maximum likelihood, artificial neural networks, and support vector machines, whereas soft approach includes Fuzzy-set classifiers. Spectral approach includes minimum distance and artificial neural networks; the contextual approach includes frequency-based contextual classifiers; and Spectral-Contextual approach includes a combination of parametric or non-parametric and contextual classifiers [147].

Miscalculation probability is one of the essential criteria for the selection of classification technique in remote sensing applications; the minimal, the better [148][149]. Supervised classification using the Maximum Likelihood Classification algorithm method (MLC) is one of the most commonly used parametric algorithms for LULC change detection as it expresses minimal miscalculation probability [150]. In this strategy, a training set is defined by identifying a homogeneous group of pixels into a set defining a class. Many such sets are created resembling different classes, and thereafter, the image pixels are classified into a class based on the similarity of their spectral responses to a given class [151].

The accuracy of the correlation between reality and the classification results can be determined with the help of an error matrix [152]. Accuracy determines the suitability of maps for drawing research inferences [153]. The minimal expected accuracy stands at 85% [154][138]. Error metrics such as overall accuracy (OA), user's accuracy (UA), producer's accuracy (PA), Error of Commission (EC), Error of Omission (EO), and Kappa Coefficient (KC) [155] estimation are critical to accuracy estimation. OA is the total number of successfully classified values over the total number of sample

values, i.e., the total number of correctly classified pixels divided by the total number of pixels. UA is the probability of a classified pixel representing an actual class real-world location. EC refers to a situation where a pixel is included in a class when it is not to be included; EO refers to a situation where a pixel is not included in a class when it is supposed to be included; PA depicts EO. KC [156] is the difference between the observed agreement and the expected agreement divided by 1- the expected agreement and ranges between -1 to 1 [157].

This study [158] concentrated on the upper Narmada River basin in central India, utilizing geospatial techniques to evaluate changes in LULC. Satellite data from Landsat 5, 7, and 8 for the years 2000, 2010, and 2022 were digitally identified utilizing the maximum likelihood technique in the ERDAS IMAGINE and ArcGIS software. This research [159] evaluates alterations in LULC and the Normalized Difference Vegetation Index (NDVI) during the past 29 years in the Wokha district of Nagaland, India, utilizing Landsat imagery. To identify land LULC changes, Supervised Classification employing the MLC method is utilized in ERDAS Imagine to categorize six primary LULC classes.

Remotely sensed data can be extensively analyzed to estimate population [37], comprehend the nature of sprawl, climate change [38], deforestation [39], urban encroachment [40], cultivation patterns [41], land use transformation, and economy. In addition to optimizing results and reducing costs, GIS enables real-time data analysis, efficient decision-making, and effective data visualization and representation. GPS is highly effective, ensures global coverage, and enables efficient and accurate localization.

These technological solutions are seamlessly integrated into research initiatives that are motivated by the study of geomorphological patterns and trends. This integration enables quick and effective analysis with higher levels of measurement and representation, thereby assisting planners and authorities in the pursuit of quality decisions for the implementation of a sustainable plan.

The efficacy of the results obtained through the use of geospatial solutions can be improved upon, verified, and validated by integrating them with ground truthing initiatives, such as social and economic investigations, censuses, and any other kind of biophysical information gathering tool, to gain a better understanding of the land use trends and the factors that influence them [42]. The nature of human activities and its impact on land use can be further understood by combining feeds from social media applications [38] and [43].

Liu et al. [44] evaluated the metropolitan level of urban sprawl worldwide by combining census statistics from 2000 and 2010 with remote sensing data. The results indicate that metropolitan areas (MAs) exhibited heightened dispersal characteristics on a global scale in 2010 when contrasted with 2000. Urban sprawl intensified as investment increased. Additionally, all urban areas expanded significantly when compared to the patterns seen during the base year of 2000 from a dynamic perspective. The results clarify the process of urban development and at the same time assist urban planners in assessing the potential risks of urban expansion and ineffective development during the different stages of urban growth. However, the analysis affirms that the multidimensional measure of sprawl provides a number of advantages for characterizing sprawl in comparison to the major measurement of density, a

conclusion that is also supported by Downs [45]. Multifaceted measurements of urban growth are advantageous due to their ability to account for both static and dynamic patterns, as opposed to solely low-density urban patterns. However, this global index is lacking.

Effat et al. employed Shannon's entropy in conjunction with multi-temporal Landsat TM and ETM imagery acquired between 1984 and 2013 to attempt to understand, identify, and quantify the spatial arrangement of Cairo's urban development in Egypt [29]. To evaluate changes in urban land-use divisions across the municipal wards and delineate built-up regions, supervised classification was implemented. The urban sprawl, trends, and spatial changes of the metropolis were modeled using Shannon's entropy. An entropy surface (index) was generated for each temporal image derived by modeling and utilizing entropy data for the city's electoral areas in an interpolation function. This index provides a visual representation of the entropy phenomena in these wards and demonstrates the spatial distribution of urban expansion. Results indicate that Shannon's entropy index increased from 1.4615 in 1984 to 2.1023 in 2013, indicating a more dispersed urban growth pattern, which is indicative of urban expansion. The demographic growth rate and areas of development were the subjects of a regression research. The results provide a visual comparison of the expansion patterns and trends within Cairo's electoral wards.

2.3.1.2 Unsupervised and Hybrid Classification

Recent advancements in LULC change detection approaches have employed remotely sensed imagery and GIS software, while several change detection methods and algorithms have been analyzed to assess their respective merits and drawbacks. The most often utilized techniques for picture classification are fuzzy classification, unsupervised classification, and supervised classification, facilitated by RS and GIS software [160][161].

We have similarly applied all the aforementioned approaches to assess the land use and cover condition of East Sikkim District, a developing district in the State of Sikkim. A variety of RS and GIS-based techniques, including supervised classification, have been utilized in the paper [162] to investigate land use/cover disruptions brought on by tourism. The same technique was also employed in the Ramnagar town region, Uttarakhand, India, to monitor changes from 1990 to 2010. Changes in LULC have drawn more attention from socioeconomic and environmental perspectives. These alterations are proliferating swiftly throughout extensive regions, particularly in emerging nations, and their environmental impact is becoming substantial [163].

Classifying multi-temporal picture data to generate thematic maps and quantify land cover alterations is a prevalent application of remote sensing. For a variety of purposes, such as land use planning, decision-making, creating land cover databases, and providing data for the sustainable management of natural resources, it is crucial to map changes in land cover at the regional level [164]. In this paper, Landsat TM pictures spanning nearly four decades were utilized. Each image was segmented in a GIS

environment and classified utilizing an unsupervised approach, subsequently followed by a supervised classification method.

A hybrid methodology was utilized to mitigate spectral confusion arising from significant land cover variability. The total accuracy of the LULC maps varied between 82.5% and 85.0%. The attained accuracies were satisfactory, and the identified classification mistakes were due to low spatial resolution and pixels including a blend of cover categories. Statistics on land cover change were retrieved and organized using ERDAS Imagine software.

In order to create a land use map, Majid et al. [165] looked at economical ways to combine Landsat images with spatial information system capabilities. This study has shown that traditional supervised and unsupervised methods produce land use maps in Iran's Golestan region that are insufficiently accurate. Decision rules based on easily accessible auxiliary data were used to synthesize an integrated methodology that combined supervised and unsupervised procedures. The accuracy of the maps produced by supervised, unsupervised, and synthetic approaches was assessed by computing confusion matrices and kappa coefficients. The synthetic method's overall accuracy was 98.2%.

Table 2.1 provide a parameterized comparison Supervised Classification and Unsupervised Classification taking into account the following parameters method, input requirement, knowledge dependency, algorithmic basis, training phase, label assignment, spectral hypothesis, computational complexity, precision potential, interpretability, adaptability to unfamiliar categories, scalability, applications, format for output, verification, and examples.

Table 2.1: A parametrized comparison between Unsupervised and Supervised Classification

Parameter	Supervised Classification	Unsupervised Classification
Method	Controlled learning utilizing annotated training data	Learning through exploration using unlabeled data
Input requirement	Ground truth or reference data is necessary.	No prior knowledge required
Knowledge dependency	A priori understanding of class attributes is crucial.	Irrespective of previous knowledge
Algorithmic basis	Discriminative estimate of class boundaries	Clustering-based (data structure discovery) or generative methods
Training Phase	Entails specific instruction utilizing sample data	Absence of a training phase; automatic grouping
Label Assignment	Pixels are allocated predetermined class labels.	Pixels are allocated to clusters, and labels are established subsequently.
Spectral Hypothesis	Presumes class-specific spectral distributions (e.g., Gaussian)	Presumes spectral uniformity within clusters

Computational Complexity	Increased, attributable to training and statistical modeling	Reduced to moderate, contingent upon the methodology and data volume.
Precision Potential	Elevated, if the training data is precise and representative	Moderate, contingent upon spectral separability and parameter adjustment
Interpretability	Direct interpretation of recognized land cover types	Demands post-classification analysis of clusters
Adaptability to Unfamiliar Categories	Subpar - constrained to recognized categories within the training dataset	Effective – capable of identifying novel or unfamiliar categories
Scalability	May incur significant processing costs for big data volumes	Frequently more scalable, yet less precise
Applications	Classification of land cover, mapping of crop types, and detection of changes (with ground truth validation)	Preliminary land cover assessment, unfamiliar terrain evaluation, data pre-processing
Format for Output	Thematic map featuring categorized classifications	Cluster map featuring numerical class identifiers

Verification	Facilitates quantitative validation (e.g., confusion matrix, Kappa coefficient)	Demands manual analysis or supplementary verification
Examples	MLC, KNN, ANN, SVM, Random Forest	ISODATA, K-Means, PCA-based clustering, Self- Organizing Maps

2.4 Data Clustering

Clustering is one of the most widely used learning-based algorithms in the field of image segmentation. It is a method of grouping data points or patterns into identifiable groups based on some predetermined criteria [166] in a manner that the data points from the same group are more similar than those from other groups [167]. Clustering is the mechanism used for grouping unlabelled data points to distinct clusters that exhibit comparative characteristics to each other than those in other clusters [159]. It can be defined as a mechanism to organize datapoints into homogeneous groups. Here, homogenous group refers to a set of datapoints that are governed by certain similarities, points that exhibit some inherent patterns or trends based on some assessment characteristics. The primary objective of clustering algorithms is to group or cluster data according to the data's innate patterns and similarities [168] and facilitate detection of anomalies/ outliers [166]. Clustering is tremendously helpful for operations such as generalizing decision on the datapoints, monitoring ease and control, and ease of statistical analysis. Numerous image-

processing based applications such as studies related to pattern recognition, machine vision, object detection, its classification and retrieval, studies related to biometric, medical imagery, and many more make use of clustering for grouping data points for ease of control and monitoring [159][169]. Several real-world application domains such as market analysis, social network analysis [168], online search, recommendation systems, information retrieval, spam email classification [166] makes extensively use of clustering techniques for efficiently grouping data into indefinable groups exhibiting very high intra-cluster similarity and very high inter-cluster dissimilarity [170]. Clustering is also extensively used in audio, image and video segmentation, content-based retrieval, big data analytics, artificial intelligence, robotics, education, marketing, medicine, biological studies, and bioinformatics [171][172]. With the emergence new application areas and onset of big data and data with higher dimensionality and complex structure clustering approaches are expected to be more adaptive and efficient [173][174].

Although clustering facilitates ease of control, it suffers from some inherent lags some of which are as follows i) the inability of the formed cluster to correlate with reality due to variation in Illumination raised due to varying lighting conditions ii) significant variations observed in the intensity of pixels belonging to same cluster and iii) presence of complex background adds to the problem due to its close resemblance with the feature of interest.

Clustering of datapoints is performed with the help of a Clustering algorithm. Based on the need of the research initiative a clustering algorithm may be conceived to create clusters of regular shapes or irregular shapes and same size or variable sizes. The

selection of clustering algorithms depends upon the purpose to be solved; this implies that the need assessment has to be comprehensive to decide on the selection of the clustering. Clustering algorithms may be broadly classified into three identifiable categories namely supervised, semi-supervised, and unsupervised depending upon reliance on the learning approaches and availability of labelled training data.

2.4.1 Unsupervised Clustering

This approach operates on unlabeled data, without being aware of the labels or categories beforehand, it aims at discovering hidden patterns or inherent structures in the data. It is a machine learning technique that helps in discovering groups of similar data points in a complex set based on closeness of the data points in the feature space using use of distance metrics such as Euclidean distance, Manhattan distance, and Cosine similarity without the need for prior training or predefined labels. The program without any prior knowledge uses only the similarity metrics to automatically find patterns or hidden structures in the data. It is an unsupervised learning technique that dynamically groups data points according to innate patterns making it often preferable for exploratory data analysis. In case of unsupervised clustering the shape, size, and the number of clusters depends on the dataset. One of the concerning issues associated with this approach is that selection of inappropriate parameters may result in creating of clusters which exhibit very poor correlation with reality, further assessing the quality of cluster is very difficult as it does not use any predefined labels. Silhouette Score, Dunn Index, or Elbow Method measures may be used to assess the quality of the clusters. Some of the popular unsupervised clustering approaches includes K-Means Clustering, Hierarchical Clustering, DBSCAN (Density-Based Spatial

Clustering of Applications with Noise), Gaussian Mixture Models (GMM), and Mean-Shift Clustering.

K-mean Because of its ease of use, effectiveness, and capacity to manage enormous datasets, the K-Means technique is frequently employed in satellite image processing. Using spectral data from satellite pictures as input, K-Means clusters effectively group pixels based on similarity of spectral signatures; each cluster represents a distinct type of land cover [175]. Some of the prominent uses of K-means includes anomaly detection (finding odd or unexpected patterns in satellite images caused by natural disasters like landslides or by human activities like mining and deforestation), creation of land cover classes (water, forest, urban areas, and agricultural land), monitoring of temporal events (change in land-cover and urbanization), analysis of vegetation using the Normalized Difference Vegetation Index to assess the health of the vegetation, cloud detection, and many more. Although k-means is simple [176], efficient and highly scalable its performance greatly depends on prior knowledge about number of clusters, and tends to underperform in situation the clusters are irregular in shape.

Hierarchical clustering [177][178] is a type of unsupervised machine learning algorithm wherein the data points are distributed among a hierarchy of clusters that advantage of this approach over k-means clustering is that it does not mandate pre-specification of number of clusters. Instead, treelike structure called are Dendrogram are used for establishing relationship between data points where the leaf nodes represent data points, merges represent the dissimilarity between clusters and branches represents cluster of data points. Hierarchical clustering techniques are of two types top-down and bottom-up called as Divisive and Agglomerative respectively. In

Divisive method initially all the datapoints are in the same cluster, and by selecting an appropriate splitting method such as DIANA (Dlvisive ANALysis) Algorithm or Hybrid Approach such as Bisecting K-means recursively the datapoints are eventually placed in their respective clusters. Whereas in Agglomerative Hierarchical Clustering Euclidean or Manhattan distance is used to calculate pairwise distances, based on the calculated distance closest clusters are merged considering linkage criterion (single linkage, complete linkage, average linkage, ward's method), further the distance is recomputed considering the new cluster and remaining clusters. The merger continues until a single cluster is left. Agglomerative clustering is relatively faster than divisive clustering. Unlike k-means clustering algorithm hierarchical clustering tends to perform well in situation when the clusters are nested and irregular in shape.

DBSCAN at creating high-density cluster by grouping points that are close to each other and further places the outliers in sparse regions [179][180]. DBSCAN algorithm consider a core point that has minimum sample points with its radius of the neighbourhood. A point is considered directly density reachable if it is within the radius of the neighbourhood of a core point. Border points are points that are relatively lesser than the sample points but are reachable from the core points. A noise point or an outlier are points that are neither a core point nor it is reachable from any core points [26]. Two points are considered to be density reachable or density connected if the points are connected via a chain of core points or a core point connected both of the them. DBSCAN algorithm does not require any prior knowledge on the number of clusters to be created and is capable of creating clusters of arbitrary size [181].

Gaussian Mixture Model (GMM) is a soft probabilistic model that aims at assigning probability to various data points belonging to different clusters assuming that points are created from mixture of Gaussian Normal Distribution [182][183]. In case of GMM each cluster is modelled as Gaussian with mean representing the centre of the cluster, covariance representing the spread of the cluster and mixing coefficient the weight of the cluster [184]. In contrast to other approaches GMM assigns probability to a data point for deciding on its inclusion to a cluster. There are two inherent steps: expectation and maximization [185][186]. Expectation step determines the probability of a point being included in a cluster whereas maximization step aims at updating Gaussian parameters: mean, covariance, and mixing coefficient for maximizing likelihood. Some of the benefits of GMM include soft approach (based on probability), capable of creating clusters of variable size, works well with overlapping features, helps on determination of density. Some of the inherent drawbacks include need for specification of number of clusters, computationally less efficient, and may generate inaccurate results in situations if the data is non-Gaussian.

Mean-Shift Clustering aims at building blob-like clusters based on centroid and density provided certain radius. It makes one fundamental assumption that the data points are samples of probability density function [187]. The bandwidth determines the radius, larger the bandwidth, lesser the clusters, smaller the bandwidth, more detailed the cluster. This approach progresses by considering every point as a potential cluster centroid. The weighted means of the points within the bandwidth is then determined and a decision is made on the shift (points to a local mean). The approach concludes when there is no significant movement ensuring maximum coverage [188]. Unlike K-means, this technique does not require any prior knowledge about the number of

clusters. It is capable of determining clusters of arbitrary shape and size. It is also capable of working with outliers as it is a density-based approach. Its performances decrease with increase in the dimensionality of data and varying density.

2.4.2 Supervised Clustering Approach

Ideally, clustering is often based on unsupervised learning framework. On the contrary, supervised clustering one of its derivatives that objectifies creation of small clusters high density with regards to close resemblance with specific characteristics of a class. It is a perfect combination of supervised learning and clustering wherein the model is pretrained using the labelled data for a guided mapping of datapoints into a cluster thereby ensuring that the cluster is coherent and meaningful in relations to the associated data labels i.e. for every data sample linked to a known cluster label, this approach offers a training dataset further the algorithm learns to allocate new data points to clusters by analysing the patterns in the training dataset [166]. This approach is extensive used in computer vision and image processing for effective object identification and localization. Supervised clustering may be based on constraints, distance metrics, or data labels. Although effective, supervised clustering is computational inefficient as it requires tremendous amount of time in situation when the dataset size itself is large, the metrics are computationally complex, when there are large number of data labels, and the nature of constraints are complex. This makes supervised clustering highly unscalable. In addition, supervised clustering mandates neat balancing of learning and clustering. Class impurity and number of clusters are the two fundamental criteria for assessing the ability of supervised clustering techniques. Some of the popular supervised clustering approaches includes Supervised Clustering using Evolutionary Computing (SCEC), Top-Down Splitting

Algorithm (TDS), Supervised t-SNE (t-Distributed Stochastic Neighbor Embedding), Supervised Partitioning Around Medoids (SPAM) and Single Representative Insertion/Deletion Steepest Decent Hill Climbing with Randomized Restart (SRIDHCR).

SCEC leverages on the capabilities of evolutionary algorithms combined with supervised learning mechanism that objectifies formation of optimal clusters [189][190] based on available labelled dataset [191] ensuring minimal misclassification. In this approach the effectiveness or the fitness [192] of clustering is determined by two critical aspect the impurity and the penalty. The impurity checks the ability of the cluster points to exhibit similarity to the class labels whereas penalty keeps tracks of optimal number of clusters. This technique hybridizes evolutionary operators like selection, crossover, etc with fitness function [193] that minimizes intra-cluster distances. It is popularly famous for handling complex high dimensional dataset with non-linear relationships.

TDS is a hierarchical clustering technique that emphasizes on recursively splitting of the dataset in smaller homogeneous clusters based on predefined splitting criteria [194]. This approach fundamentally relies on two steps initialize and split that are recursively executed. Some of the commonly used splitting techniques are density-based splitting, bisecting K-means, and Principal Direction Divisive Partitioning (PDDP) re During splitting, the objective of the method is to maximize inter-cluster dissimilarity or to minimizing intra-cluster variance. This technique popularly used in situation when the size of the dataset is large and demands global optimization. However, the performance of the algorithm relies on splitting criteria [195].

It is a supervised version of t-Distributed Stochastic Neighbour Embedding algorithm [196] (t-SNE) that makes use of labelled data in the embedding process to create cluster present in high-dimensional space. In contrast to t-SNE [197] which aims at preserving only structure, supervised aim at preserving both the structure and separation. During transformation it preserves the pairwise similarity, thus, the cohesiveness of formed clusters is high, stable and reliable while maintaining the clear separation between different class. Supervised -SNE efficiently separates labelled clusters, and it is preferred in situation where not only class separations but class structure is also prioritized.

Supervised Partitioning Around Medoids (SPAM) [198] is a supervised version of the traditional Partitioning Around Medoids (PAM), that leverages on the capabilities of labelled data or domain knowledge in assisting the clustering process in formation of compact and well-separated clusters that has better meaning and are more interpretable. Unlike traditional PAM (unsupervised) which assigns priority to only features, SPAM [199] uses dissimilarity measures like Euclidean distance, SPAM (Supervised) modifies the objective function that uses known data labels that directs to form cluster with same class. SPAM is highly flexible as it can work with both labels or constraints, it facilitates interoperability as it relies on medoids, and finally it can efficiently handle outliers and noise making it extremely robust.

SRIDHCR [200] makes use of Hill Climbing optimization and random restart to determining optimal clusters. This technique uses greedy optimization approach to avoid local optima problem to create a robust solution by restarting from a new random configuration. For improving the performance of clustering process, the technique uses

steepest descent hill climbing approach where at every iteration the technique evaluates all possible insertion and deletion of centroid. Single representative ensures each cluster is represented by single cluster head. Mix of different strategies makes technique robust and computationally intensive. This technique is highly flexible, and adaptive.

2.4.3 Semi-supervised Clustering

This approach leverages on the capabilities of both supervised and unsupervised learning for improving the performance by providing class labels and associated constraints that suitably governs the working of clustering approaches. It relies on labelled and unlabelled data values for directing the learning process to facilitate optimal partitioning of the dataset. The overall objective of semi-supervised clustering is to create a flexible framework for improve accuracy, qualisty, scalability, and performance by appropriately deciding on the constraints. On the contrary this approach may be also prone to the negative influence of inappropriate constraints, unbalanced representation of labelled and unlabelled data, as well as extensive computation requirements. Some of the popular Semi-supervised clustering approaches includes Constrained K-Means, Constrained Gaussian Mixture Models (GMM), Semi-Supervised Support Vector Machine (S3VM), Constrained Spectral Clustering, and Seeded K-Means,

Constrained K-Means algorithm is popular semi-supervised extended K-means algorithm that works on the principle of clustering data points based on user-defined constraint which includes must-link constraints, cannot-link constraints and the size constraints [200]. The constraints help in enforcement of domain knowledge during

cluster formation. Thus, the integration of constraint aids in minimal intra-cluster variance with improved accuracy. This approach is preferred in situation when there is very limited knowledge available before the onset of clustering process [201].

Constrained Gaussian Mixture Models (GMM) The performance of traditional GMM is poor in a situation where the dimension of data is high or structured [202][203]. A guided version of GMM referred to as Constrained GMM provides greater flexibility by incorporating prior knowledge through constraint during model building. The developed model ensures that learned parameters adhere to specified constraints [204][205]. The constraints are imposed during Expectation-Maximization (EM) optimization process. The model performs well for higher dimensional data with improved interpretability and reduced overfitting situation where there exists a partially labelled dataset.

Semi-Supervised Support Vector Machine (S3VM) or S3VMs [206] is an extension of traditional SVM which uses both labelled and unlabelled data to perform classification. Unlike traditional SVM, where model is built to maximize the margin between labelled data of different class [207][208], the S3VM takes the advantage of unlabelled data to maximize the margin of decision boundary. During optimization of optimal hyperplane, S3VM tries to minimize the classification error on labelled data while maximize the margin for unlabelled data using commonly used optimization techniques like gradient descent, etc. S3VM is popularly used in situations where obtaining label for data is expensive and time consuming. Since S3VM depends on labelled data to aid the learning process for unlabelled data, hence, it is sensitive to initial labelled data. Thus, the initial labelled data must be reliable.

Constrained Spectral Clustering [209][210] is built over traditional Spectral Clustering algorithm that ensure implementation of constraint or prior knowledge during cluster formation. The clusters formed using constraint aims to generate more efficient partitioning of dataset as compared to unsupervised technique. The algorithm takes the leverage of few labelled data to aid unlabelled data keeping into consideration the constraints conceived for cluster formation. However, the performance of the model totally relies on the quality and consistency of the defined constraints. In addition, the noise may also influence the quality of outcome.

Seeded K-Means [211] is a semi-supervised algorithm that takes help of few labelled data, called seed, that will guide the clustering process. These seed act as initial cluster head for cluster formation. Using seed information, it avoids local optima problem thereby improving curacy. However, the selection of seed plays crucial role in performance of the model. If the seed chosen is not reliable or noisy then the reliability of the model is either less or lost.

Table 2.2: Inferences drawn from study of clustering techniques

Classification	Clustering	Type	Cluster Shape	Complexity	Number of clusters (K)	Ability to handle Noise	Advantages	Disadvantages	When to use?
Unsupervised	K-Means	Centroid based	Spherical	$O(nkt)$	Yes	Weak	- Fast and Efficient. - Creates spherical clusters. - Converges to local optimum.	- K needs to be predefined. - Creates only spherical clusters. - May Create non-optimal clusters. - Works poor with noisy data.	- When dataset is large and clean. - When dataset expresses spherical clusters.
	Hierarchical	Agglomerative or Divisive	Arbitrary	$O(n^3)$	No/ Yes	Average	- K need not be predefined. - Works with any type of metric. - Creates arbitrary shaped clusters.	- Takes greater amount of time. - Sensitive to outliers and noise. - Increased memory utilization.	- When dataset is of relatively smaller size. - When dataset inherently expresses hierarchy.

								Result depends on the preference of linkage method.	
	DBSCAN	Density-based	Arbitrary	$O(n \log n)$	No	Very Strong	- K need not be predefined. - Creates arbitrary shaped clusters. - Effectively handles noise and outliers.	- Takes considerably more time. - Sensitive to hyperparameters. - Dimensionality influences efficiency. - Suffers from border point anomaly.	- Suitable for spatial data. - Noisy dataset. - Dataset expresses clusters of arbitrary shape.
	GMM	Probabilistic	Ellipsoidal	$O(nkd^3)$	Yes	Average	- Handles uncertainty. - Creates arbitrary shaped clusters. - Works well with overlapping clusters. - Efficiently handles noise.	- Underperforms when the clusters are non-Gaussian. - K needs to be predefined. - Dimensionality influences efficiency. - Computationally inefficient.	Suitable for overlapping cluster.

	Mean-Shift	Density-based	Blobs	$O(n^2)$	No	Strong	• K need not be predefined. • Adaptive approach allowing non-linear and irregular shaped clusters. • Effectively handles outlier. • Does not consider data distribution.	• Computationally expensive. • Kernel dictates number of clusters. • Dimensionality influences efficiency. • Increased reliance on density. • Consumes greater computational memory.	• Suitable in situation if the number of clusters are not known. • Clusters are expected to be irregular and of variable sized. • Dataset has outliers.
Supervised	SCEC	Clustering (Supervised, Evolutionary)	Non-linear, arbitrary shapes	$O(p \cdot k \cdot n(d + \log c))$	Optimized via evolution	High (Fitness functions filter noise)	• This technique is highly adaptive and handle complex data • Creates optimal number of clusters • Determines fitness based on impurities (misclassification) and	• Computationally expensive • Suffers from overfitting • The effectiveness of the technique greatly depends on the fitness function	• Recommended for high-dimensional or non-linear data with labels and noisy data, where accuracy is critical

							penalty (optimal number of clusters)		
	TDS	Hierarchical Clustering (Divisive)	Tree-like Structure (can model	$O(n \log n)$	Automatic (Based on Splitting/Stopping Criteria)	Low (Splits may amplify noise)	• Computationally efficient • Facilitates effective data interpretation • Highly adaptive • Flexible and scalable for handling larger data	• Highly sensitive to splitting criteria • May result in creation of sub optimal clusters that are non-re-adjustable	• Recommended for operations where establishment of hierarchy among data points is important
	Supervised t-SNE	Dimensionality Reduction (Supervised)	Non-linear, arbitrary	$O(n^2)$	Predefined or post-visualization	Moderate (Sensitive to outliers)	• Preserves class structure in 2D/3D • Good for visualization	• Computationally heavy	• Visualizing high-dimensional labeled data

Semi -Supervised	SRIDHCR	Optimization/Clustering (Heuristic)	Varies (depends on objective)	High (due to iterative)	Yes (Fixed/ variable)	Moderate (Restarts help)	- Avoids local optima - Precise refinements	- Slow convergence - High computational cost	For complex clustering problems with many local optima
Constrained K-Means	Clustering	Spherical	$O(n * k * d * t)$	Fixed (k)	Moderate	Simple and efficient; Improves accuracy with domain constraints; reduces intra-scluster variance	Constraints must be well-defined; sensitive to constraint quality	When prior knowledge is available. Also, if the dataset is partially labelled and need fast clustering	
	SPAM	Clustering (Supervised)	Spherical (Medoid-	$O(k(n-k)^2)$ - Moderate to	Yes- Fixed	Robust (Medoids reduce noise impact)	- Robust to noise - Interpretable (medoids)	- Struggles with non-spherical clusters - K must be known	When labeled data is available and representative points are needed

	Constrained GMM	Probabilistic	Elliptical	$O(n * k * d^2 * t)$	Fixed (k)	Moderate	Handles high-dimensional data; reduces overfitting; Improved interpretability. Handles noisy data.	Slower than K-Means; sensitive to constraint violations; Computationally expensive; may overfit on limited dataset.	For higher dimensional or structured data with partial labels; can be used when dataset has noise.
	Semi-Supervised SVM (S3VM)	Classification	NA	$O(n^2 \text{ to } n^3)$	NA	Low	Leverages unlabeled data for better generalization; Less expensive and less time consuming.	Sensitive to initial labeled data; assumes cluster structure; prone to noise.	When labeled data is limited but unlabeled data is abundant
	Constrained Spectral Clustering	Graph-based clustering	Arbitrary	$O(n^3)$	Fixed (k)	Low	Works well with non-linear data; better partitioning; Implements constraint strictly.	Computationally expensive; sensitive to constraints. Very prone to noise.	For graph data with pairwise constraints; Useful when limited labelled data are available.
	Seeded K-Means Clustering	Spherical	$O(n * k * d * t)$	Fixed (k)	Low	Avoids poor local optima; more stable initialization; Resulting	Performance depends entirely on seed quality; Prone to noisy seed.	When reliable seed labels are available and with limited labelled data	

							cluster is stable and accurate.		
--	--	--	--	--	--	--	---------------------------------	--	--

2.5 Identification of critical factors influencing urban sprawl

The research initiative also aims at identifying, analysing and determining trends of a set of critical factors that may be influential in the growth of urban sprawl which includes:

a) Population details from the year 1981 to 2022 [212][213][214][215][216] is presented in table 2.3.

Table 2.3: Description of the Population Distribution

Description	1981	1991	2001	2011	2022
Urban Sikkim	51084	36984	59870	153578	164432
Urban East Sikkim	43242	31821	52852	122487	131201
Urban Rest of Sikkim	7842	5163	7018	31091	33231
Rural Sikkim	265301	368.521	480981	456999	473618
Total Population	316385	405505	540851	610577	638050

b) Number of households from the year 1981 to 2022 [212][213][214][215][216] is presented in table 2.4.

Table 2.4: Description of the Number of Households

Household	1981	1991	2001	2011	2022
Urban East Sikkim	9063	10334	11605	28691	39719
Urban Rest of Sikkim	1603	1498	1393	7027	8985
Urban Sikkim	10666	11832	12998	35718	48704
Rural East Sikkim	18218	30597	42976	32876	41364
Rural Rest of Sikkim	30973	44611	58249	60412	72005
Rural Sikkim	49191	75208	101225	93288	113369
Total East Sikkim	27281	40931	54581	61567	81083
Total Rest of Sikkim	32576	46109	59642	67439	80990
Total Sikkim	59857	87040	114223	129006	162073

c) States Institutional Administrative Units [217]

- There are 46+ departments under the state government of Sikkim, all of the 46 departments have their headquarters in the East Sikkim District.

d) Academic Institutions [218][219]

- There are 1397+ schools in the State of Sikkim out of which 472+ schools are in East District.
- There are 23+ colleges in the State of Sikkim out of which 16+ colleges are in the East District.
- There are 17+ universities in the State of Sikkim out of which 7+ premiere universities are in the east district.

e) Health Institutions [220]

- There are 184+ health institutions in the State of Sikkim out of which 60+ institutions are in East Sikkim.

f) Tourism Entities [221]

- There are 768+ hotels in the State of Sikkim out of which 613+ are in the East District.
- There are 1305+ homestays in the State of Sikkim out of which 597+ are in the East District.
- There are 30+ tourist destinations in the East District.

g) Infrastructural Projects

- There are 34+ power projects in the State of Sikkim under Independent Power Producer (IPP), Joint venture and Sikkim Power Development Corporation Limited (SPDCL) out of which 10+ power projects are in the East District [222].
- There are more than 10+ government projects in the State of Sikkim out of which 7+ are in the East District.

f) Industries and Manufacturing units [223][224][225]

- There are 28+ Pharmaceutical Companies in the State of Sikkim out of which 21+ are in East District.
- There are 8+ Distilleries in the State of Sikkim out of which 4+ are in East District.
- There are 2+ Breweries in the State of Sikkim out of which 1 is in East District.
- There are 10+ manufacturing industries in the State of Sikkim out of which 7+ is in East District.

2.6 Data Prediction Model

Data prediction models are mathematical models use to predict the future datapoints considering the past trends expressed by the datapoints. Categorically these models may be statistical or non-statistical in nature. Statistical models are based on mathematical formulation such as probability, where as non-statistical models are based on determination of patterns evident in the data. Statistical models are highly interpretable, easy to use and applicable on small dataset but is extremely constrained where as non-statistical models are highly complex, and applicable to large dataset but extremely flexible. Some of the examples of statistical model are Linear regression, Auto Regressive Integrated Moving Average (ARIMA). Whereas Random Forest and Agent-Based Modeling (ABM) are examples of non-statistical model [226].

2.7 Inference from Literature Survey

- Urban sprawl is the low-density, haphazard, uncontrolled, unplanned, and unrestricted spread of urban spaces into barren, green, and farm lands as a consequence of accelerated developmental initiatives.
- Urban sprawl can be broadly classified into two broad categories: radial and linear. A radial sprawl is observed spreading out from an urban center, expanding outwards towards the peripheries, whereas a linear sprawl is observed along a linear feature such as a road network or a river network, etc.
- Some of the specific types of urban sprawl include leapfrog sprawl, ribbon sprawl (also known as strip sprawl), low-density single-use sprawl, exurban, and edge cities.

- Remotely sensed satellite images can be used as a scientific basis for understanding, studying, analyzing, and conceiving mitigation plans for urban sprawl.
- One such technique is image classification, where a remotely sensed satellite image can be processed to detect land use and land cover.
- Supervised, unsupervised, and self-supervised image classification-based techniques are the potential candidates for selection.
- It has been observed that supervised classification-based techniques are mostly preferred in such situations due to their efficiency in terms of their time requirement, performance accuracy, and execution ease. Unsupervised approaches do prove effective in situations where the basis for classification is not explicit but consume more time. However, the capability of both supervised and unsupervised classification techniques can be leveraged for verifying improvement in accuracy.
- However, the performance of these three approaches on a given dataset may be determined to check for efficiency.
- Performance in terms of accuracy can be estimated by determining user's accuracy, producer's accuracy, error of omission, error of commission, kappa coefficient, and overall accuracy.
- Determining trends in urban sprawl mandates the availability of temporal satellite images. The correctness of the results from the classification techniques is highly subject to the quality of input images. Therefore, it is necessary to collect clear images.
- Signatures representing various classes may be set as an input to the selected classification technique for processing.

- Sequentially process all the temporal satellite imagery and the results obtained may be statistically analyzed to build an understanding of the evident trends in the land use and land cover patterns with a focus on urban sprawl.
- On achieving desired accuracy from the classification process, the class representing urban sprawl may be selectively placed to a distinct layer for further processing.
- A set of well-defined seed points representing the urban cores may be selected for determining the temporal changes in the density of the clusters.
- The selection of seed points should be done considering three crucial facts that include the ability of a seed point to a) create high-density clusters, b) ensure maximum coverage, and c) maintain minimum distance from another seed point.
- The clusters thus formed around a seed point across the temporal satellite images may be statistically analyzed to understand the growth trends.
- Further to support the confidence of the results obtained, other critical factors may also be accounted for, such as population, households, and other infrastructural developments such as administration, academics, business centers, health facilities, industries, and government projects.
- Visualization tools may be further used to express the inferences.
- Thereafter, predictive analysis may be performed to forecast possible future trends.