Learning and Reasoning Multifaceted and Longitudinal Data for Poverty Estimates and Livelihood Capabilities of Lagged Regions in Rural India

Atharva Kulkarni[1] , **Raya Das**[2] , **Ravi S. Srivastava**[3] and **Tanmoy Chakraborty**[4]
[1]Language Technologies Institute, Carnegie Mellon University, USA
[2]Indian Council for Research on International Economic Relations, New Delhi, India
[3]Centre for Employment Studies, Institute for Human Development, Delhi, India
[4]Indian Institute of Technology Delhi, India
atharvak@cs.cmu.edu, rdas@icrier.res.in, ravi.srivastava@ihdindia.org, tanchak@iitd.ac.in

Abstract

Poverty is a multifaceted phenomenon linked to the lack of capabilities of households to earn a sustainable livelihood, increasingly being assessed using multidimensional indicators. Its spatial pattern depends on social, economic, political, and regional variables. Artificial intelligence has shown immense scope in analyzing the complexities and nuances of poverty. The proposed project aims to examine the poverty situation of rural India for the period of 1990-2022 based on the quality of life and livelihood indicators. The districts will be classified into 'advanced', 'catching up', 'falling behind', and 'lagged' regions. The project proposes to integrate multiple data sources, including conventional national-level large sample household surveys, census surveys, and proxy variables like daytime, and nighttime data from satellite images, and communication networks, to name a few, to provide a comprehensive view of poverty at the district level. The project also intends to examine causation and longitudinal analysis to examine the reasons for poverty. Poverty and inequality could be widening in developing countries due to demographic and growth-agglomerating policies. Therefore, targeting the lagging regions and the vulnerable population is essential to eradicate poverty and improve the quality of life to achieve the goal of 'zero poverty'. Thus, the study also focuses on the districts with a higher share of the marginal section of the population compared to the national average to trace the performance of development indicators and their association with poverty in these regions.

1 Introduction

Poverty is a complex situation in which the lack of capabilities translates to low income of the household [Nussbaum and Sen, 1993]. From a monetary perspective, poverty can be described as an interlacement of income distribution below a threshold value and the disproportions that exist within that boundary [Balaji, 2020]. It is a state of destitution in which individuals lack the basic essential means, such as food, water, shelter, and money, required to sustain their daily livelihood [A.M, 2020]. While poverty has been a perennial socio-economic problem of mankind, the last two decades have witnessed a steady decline in global poverty [Bank, 2020]. However, owing to the COVID-19 pandemic, the compounding effects of climate change and socio-economic conflicts, the pursuit to end poverty has suffered a significant setback for the first time in a generation. The pandemic resulted in the addition of almost 100 million people diving into extreme poverty [Bank, 2020]. Moreover, as the United Nations (UN) lists poverty eradication as one of its primary Sustainable Development Goals (SDGs), global communities are striving hard to develop efficient techniques for poverty tracking, estimation, and eradication.

India under crisis: As per Agriculture Census, 2015-2016[1], around 68% of the population resides in rural India. The latest Multidimensional Poverty Index (MPI) scores indicate that the poverty score is 32.75% among the rural population, contrary to 8.8% in urban India. The BIMARU[2] states continued to have the most deprived districts of the country, with some statistics being comparable to Sub-Saharan African countries, thus, calling into question the inclusiveness of policies in India. Around 50% population of the country is engaged in agriculture and allied sector; therefore, the regional development of agriculture has an immense impact on the quality of life of rural households. Indian society has a hierarchy with different levels of mobility across social groups. As per the NFHS-4 data, two minority classes, namely scheduled caste (SC) and scheduled tribe (ST) households, have more poverty prevalence than general social groups.

Poverty estimation is crucial: For any nation, accurately measuring poverty statistics and the economic characteristics of the population critically influences its research and national policies [Škare and Družeta, 2016]. Moreover, economic growth cannot be the exclusive goal of a nation's economic policies; it is equally vital to ensure that the benefits

[1] https://agcensus.nic.in/document/agcen1516/T1_ac_2015_16.pdf

[2] BIMARU is an acronym formed from the first letters of the names of the Indian states of Bihar, Madhya Pradesh, Rajasthan, and Uttar Pradesh.

of economic prosperity reach all segments of society. This underpins the importance of assessing poverty in all its manifestations. Also, poverty measurement is crucial to evaluate how an economy is performing in terms of providing a certain minimum standard of living to all its citizens. In summary, measuring poverty has significant implications for policy drafting and implementation. As a result, it is no surprise that poverty estimation and tracking have garnered the attention of economists, social scientists, and statisticians alike.

Multifaceted measurement of poverty: Poverty, however, is not just based on the monetary distribution of wealth amongst the masses but is a multifaceted idea comprehensive of social, financial, and political components. Thus, to unthread the fabric of poverty and understand its nuances, a deeper and multidimensional study of its different facets is required. Such *multidimensional measurement of poverty* encompasses two approaches – poverty as capability deprivation, and poverty as a measure of deprivation [Atkinson, 2003]. The Multidimensional Poverty Index (MPI), jointly developed by the Oxford Poverty and Human Development Initiative (OPHI) and United Nations Development Programme, considers both these factors (incidence and intensity of deprivation) for measuring poverty. The widely adopted Alkire and Foster's methodology [Alkire and Foster, 2011] considers the three indicators of standard of living, education, and income at the household levels to measure multidimensional poverty. However, in the case of developing countries, such as India, one must look beyond these aspects, as here, poverty estimation is beset by several quantitative and qualitative concerns. This calls for the consideration of other ancillary factors, such as the variance in climate, healthcare facilities, dietary habits, political status, cultural influence, infrastructure development, and geographical benefits and drawbacks along with the financial information [Ahluwalia *et al.*, 1980].

Existing studies on India-specific poverty estimation and research gaps: For India, the traditional narrative on poverty estimation makes use of publicly-available data collected through household surveys, such as the National Family and Health Survey (NFHS) [Mohanty, 2011; Chaudhuri *et al.*, 2013; Mishra and Ray, 2013], National Sample Survey (NSS) [Sarkar, 2012; Mishra and Ray, 2013], Indian Human Development Survey (IHDS) [Dehury and Mohanty, 2015] to name a few, to calculate the multidimensional poverty index. The census survey of the government of India publishes village-level statistics, and ICRISAT publishes the longitudinal village-level data on land relations, crop yield, access to facilities, and resource possession. However, due to the paucity and the lack of comparability, veracity, and precision of the data at different time periods, accurate estimation and tracking of poverty are challenging. Thus, recent studies in the realm of poverty estimation have focused on alternate and proxy data sources, such as satellite images [Jean *et al.*, 2016; Xie *et al.*, 2016; Ayush *et al.*, 2020; Ayush *et al.*, 2021], mobile phone usage [Smith-Clarke *et al.*, 2014; Blumenstock *et al.*, 2015; Smith-Clarke, 2021], geospatial information [Puttanapong *et al.*, 2020], and information on the web [Sheehan *et al.*, 2019]. Along with being inexpensive to produce and easier to scale, the use of these data sources also introduces the scope for utilizing modern machine learning (ML) techniques for poverty and quality of life estimation. While the use of artificial intelligence (AI) is a regular feature for achieving other sustainable development goals, such as climate change, education, healthcare, clean energy, and gender equality, its application to poverty tracking has been limited. Moreover, studies on poverty estimation in India have exclusively focused on household survey data, excluding the possibility of using alternative proxy data sources.

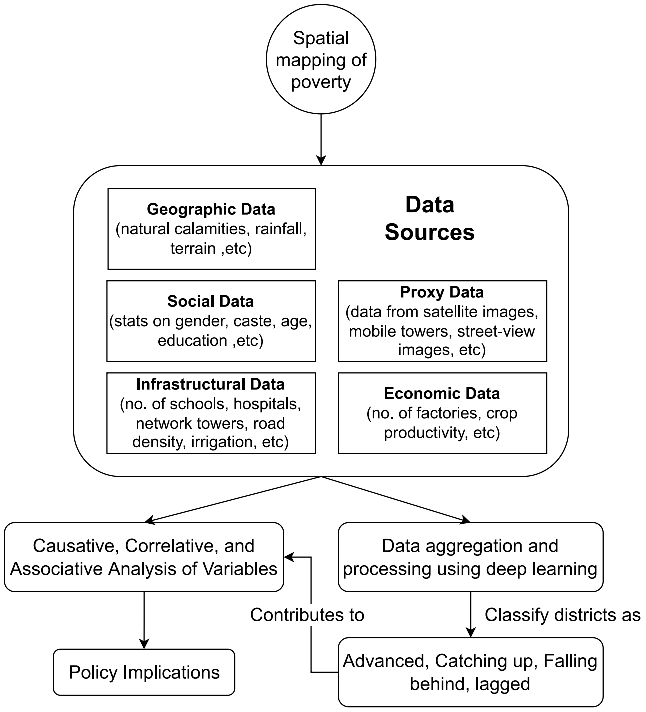
Spatial mapping of poverty
Data Sources
Geographic Data (natural calamities, rainfall, terrain ,etc)
Social Data (stats on gender, caste, age, education ,etc)
Proxy Data (data from satellite images, mobile towers, street-view images, etc)
Infrastructural Data (no. of schools, hospitals, network towers, road density, irrigation, etc)
Economic Data (no. of factories, crop productivity, etc)
Causative, Correlative, and Associative Analysis of Variables
Data aggregation and processing using deep learning
Contributes to
Classify districts as
Policy Implications
Advanced, Catching up, Falling behind, lagged

Figure 1: A bird's-eye view of our proposal.

Longitudinal and temporal analysis of poverty estimate: While India has witnessed a commendable rise in its economic growth post-independence, it has been non-inclusive and exclusionary, exacerbating the rural-urban divide, regional disparities, and social and gender-based inequalities [Dev, 2010]. Studies on analyzing multidimensional poverty in India highlight the disparity at regional levels, where eastern and central states reflect high MPI scores [Das *et al.*, 2021]. The states of Jharkhand, Uttar Pradesh, Rajasthan, Odisha, Bihar, Chhattisgarh, and Arunachal Pradesh, have a higher poverty head-count ratio while Kerala, Mizoram, Nagaland, Punjab, Himachal Pradesh, and Haryana have lower poverty rate [Tripathi and Yenneti, 2020]. Moreover, the temporal distribution of poverty has not been constant, as, over time, the poverty has aggrandized in the states of Manipur and Arunachal Pradesh in the rural setting, and Meghalaya, Odisha, and Jharkhand in the urban setting [Tripathi and Yenneti, 2020]. This underlines the need for temporal and longitudinal analysis of poverty estimation. With the amalgamation of traditional as well as proxy data sources, such investigations will help unravel the peculiarities of areas diving south of the poverty line, thus, enabling the policymakers to establish suitable and appropriate poverty eradication guidelines. Similarly, on other hand, identifying the regions where the poverty index has increased over time will help diagnose the policies that worked and the ones which did not. Lastly, it will also help differentiate the geographical regions of chronic poverty from transient poverty.

Major contributions: The proposed study investigates the spatial mapping of poverty by the quality of life and livelihood capabilities in the lagged regions vs. advanced regions of rural India at the district level. We also intend to a *longitudinal and temporal analysis* to focus on the trajectory of growth and catching up capabilities of lagged regions of the country. We further herald a new research direction towards estimating poverty in lagged Indian states using data aggregation and integration and expound on its application using ML. Figure 1 shows a bird's eye view of our overall proposal. Thus, with the poverty-struck Indian states as a case study, we aim to address the following broad research questions:

1. **Multidimensional data integration:** How different types of proxy data sources be used to estimate poverty in India?
2. **Efficient prediction:** How can these proxy data sources be integrated with the traditional survey data for more accurate, interpretable, and efficient poverty estimation through traditional ML and advanced neural models?
3. **Longitudinal analysis and estimation:** Can a temporal and longitudinal examination of poverty using proxy data sources reveal salient information about the factors that contribute to poverty trajectory?
4. **Correlation, association, and causality analysis:** Can we identify variables that have a direct causal, correlative, and associative effect on poverty estimation?
5. **Policy Implications:** How can causal, correlative, and associative analysis, along with the classification and forecasting study help in policymaking?

2 Goals

The primary purpose of this research is to demonstrate the shortcomings of the conventional approach to estimating poverty in India and to illustrate new approaches and methodologies for doing so with the recent advances in AI and ML. Furthermore, this study emphasises the need for a temporal and longitudinal analysis of poverty, rather than a static and time-constrained one, in order to fully comprehend the expression and expansion of poverty in the socio-economically backward states of India. This section expands on these goals.

2.1 Limitations of Traditional Data Sources

The data from the household expenditure and income surveys, such as the National Family and Health Survey (NFHS), different rounds of the National Sample Survey (NSS), and the Indian Human Development Survey (IHDS), form the backbone for identifying and measuring the poverty status of Indian households. Although the census data provides a comprehensive measurement of the material living standard of individuals in a population, it is conducted over long cycles (usually every ten years) and encompasses only a few characteristics of the target population. Household surveys, on the other hand, cover a wide range of variables, but their reliability for local statistical inference is limited. Conducting such surveys is also expensive, time-consuming, imprecise, and at times infeasible for poverty assessment. The majority of these statistics are generated over a reasonably long period and do not accurately reflect the attributes that affect the livelihoods of the residents of a particular area. Moreover, the target population is also not the same for all the datasets. For example, the states of Arunachal Pradesh, Bihar, Jharkhand, and Orrisa have a considerable tribal population along with the standard rural and urban distribution. This raises concerns over the credibility of the estimated poverty figures. With a few exceptions, such as information about education and job status, such surveys focus solely on household monetary information [Alkire and Foster, 2011]. However, as poverty is not a uni-dimensional phenomenon and not just a mere reflection of the economic status of an individual, it encompasses other concomitant aspects, such as data on infrastructural development, agricultural growth, vehicle count, network towers, political climate, caste information, electricity usage, and so forth. Such elements are more relevant in rural settings as they are direct markers of prosperity. While household data comprising the income statistics does have its place in poverty prediction, they have certain limitations, as stated above. Therefore, *we propose a data integration mechanism that takes into account both the traditional datasets along with the new proxy data sources for poverty prediction*.

2.2 Towards the Use of Proxy Data Sources

There have been numerous studies on multidimensional poverty estimation, albeit most of them incorporate just the conditions of education, health, and living standards in addition to the economic dimension [Alkire and Foster, 2011]. While these strategies are unquestionably superior to the single-source poverty assessment methodologies, they do have some detriments, as discussed in the previous section. To compensate for these deficiencies, the advent of big data, combined with technological breakthroughs in ML, offers great promise for poverty tracking as well as interpreting and predicting social-economic conditions. This accounts for the data gathered from social media, remote sensing, agricultural growth, vehicular traffic, infrastructural development, mobile phone meta-data, and housing details. Political, cultural, and environmental aspects also factor in while estimating poverty. There are examples aplenty that elaborate on the successful use of these proxy data sources for poverty estimation across geographies. For instance, remotely sensed images, such as Landsat data and night-time light images, serve as the most representative data sources as they provide important information about the region's landscape. Various case studies on China [niu, 2020; Shi *et al.*, 2020], Thailand [Puttanapong *et al.*, 2020], Philippines [Hofer *et al.*, 2020], Bangladesh [Steele *et al.*, 2017], and African countries [Jean *et al.*, 2016; Ayush *et al.*, 2021] have demonstrated the usability of remote sensing for poverty tracking and estimation. Some other studies have shown that employing data from mobile phones [Blumenstock *et al.*, 2015], communication networks [Smith-

Clarke *et al.*, 2014], and political climate [Van der Berg *et al.*, 2006] also yields promising results. While these research works do validate the use of proxy data sources for poverty estimation, they also bring to light that no such study has scrutinized the Indian subcontinent. Moreover, these works use independent proxy data sources for poverty estimation and do not contemplate the idea of combining and integrating different data sources for more efficient, explainable, and robust predictions. Thus, this works aims to address this bottleneck by proposing a data aggregation and integration methodology, combining the traditional as well as proxy data sources, for poverty estimation of Indian states. In addition, we offer a multi-input deep learning-based architecture for aggregating and processing data from many sources. Finally, we provide methods for investigating the correlative, associative, and causative relationships between various input variables and poverty estimation, the identification of which can help in better policymaking.

2.3 Temporal Analysis of Poverty

Poverty is a temporal phenomenon that aggravates or declines over time. Thus, to forecast poverty statistics and its progress in a region, its historical data characterizing the temporal shifts should also be taken into consideration. While such a methodology would, in all likelihood, yield better, more accurate, and robust predictions, it will also help differentiate the prominent factors contributing towards poverty from the weaker ones. It will also help diagnose the schemes, plans, and policies that succeeded during a specific time leading to poverty alleviation or that had an adverse effect, leading to poverty expansion. The contemporary literature on poverty estimation exclusively focuses on static and time-specific data, failing to account for the temporal aspect of poverty. In rapidly developing countries like India, it is paramount to take into account the temporal dimension of poverty to pinpoint its core causes and design policies to tackle it. Thus, this research proposes a temporal data collection, integration, and prediction scheme for more robust poverty forecasting.

3 Methodology

In this section, we elaborate on the different proxy data sources, the data integration methodology, the use of ML and deep learning techniques for poverty estimation, and the temporal analysis of poverty. Furthermore, we classify the districts based on the performance (MPI score) between 1993-2015 considering several rounds of NFHS household surveys into ‘advanced’, ‘catching up’, ‘falling behind’, and ‘lagged’ districts. The study further targets the four most lagged states – Uttar Pradesh, Jharkhand, Bihar, and Odisha, as they rank the highest in the multidimensional poverty index of 2021 [Tripathi and Yenneti, 2020].

3.1 Proxy Data Sources and Data Integration

To develop a functional methodology that governments of developing countries can use for accurate poverty estimation and tracking, one requires a dataset that is representative of the country’s population, that can be collected and timely updated automatically, and that is available at a fine level geographical granularity. In this section, we explore the different proxy data sources for poverty estimation as well as elaborate on the data integration approach.

Remote Sensing

Remote surveying with satellite imagery is a low-cost and dependable method of tracking human development at fine spatial and temporal resolution [Jean *et al.*, 2016]. Remote sensing involves various types of satellite imaging, conveying different types of information during the day and night. Daytime satellite images provide a wealth of information on the region’s geography, infrastructure development, and population growth. Moreover, it can be used to infer other signals of prosperity, such as growth in the road network, building density, forest cover, and infrastructural expansion. The nighttime luminosity information provided by the satellite images provides a lens over a region’s nocturnal activity. It serves as an ideal proxy for electricity consumption, degree of electrification and population growth [Ghosh *et al.*, 2013]. Studies also showed a positive correlation of nighttime luminosity with carbon dioxide emissions, GDP, GDP per capita, constant price GDP, non-agricultural GDP, and capital stocks [Addison and Stewart, 2015]. Therefore, we aim to collect and utilize both daytime and nighttime satellite image data as each has its own advantages. We propose to use the satellite imagery from the Landsat 7 mission from the years 2001, 2011, 2016, and 2019 to track daytime activity. The nighttime light data can be procured from the United States Air Force Defense Meteorological Satellite Program (DMSP). The satellite’s Operational Linescan System (OLS) sensors have a spatial resolution that allows them to make observations ranging from entire continents to less than a square kilometre. To map villages and districts to their satellite-image locations, we also propose to collect the information on each of their *shapefiles*. Once a region has been linked to its satellite images, we can extract its human development attributes from the satellite-image features for that region, as stated above, based on its daytime and nighttime visual look from space.

Communication Networks

Active research in the last decade has shown that information from mobile phone usage and telecommunications networks are a strong indication of a region’s socio-economic status [Soto *et al.*, 2011; Smith-Clarke *et al.*, 2014; Pokhriyal *et al.*, 2015]. We can automatically infer proxy indicators of poverty from unobtrusively obtained call network data by a detailed examination of patterns inherent in mobile phone users’ collective behaviour. For instance, cellphone top-up behaviour suggests that poorer people are likely to top-up their phone credit regularly in small amounts, whereas wealthier people are more likely to top-up infrequently in larger amounts. Furthermore, an increase in network density of communication reflects an infrastructural development, an increase in population, and socio-economic well-being. In our study, *we propose to collect the Call Detail Records* (CDR) *data aggregated to the cell tower level.* The mobile phone operators collect such user-specific data primarily for billing purposes. Using such data, we can obtain a wealth of information about each call or text message, including the time, duration, caller and callee IDs, as well as the base station towers routing the call or text. Such precise data not only reveal the degree of the penetration rate of mobile technology in developing countries but also provides a relatively unbiased picture in terms of demographics. To protect users' privacy, we plan to collect the CDR data aggregated by the cell towers through which the calls are routed rather than using the data at an individual user level. *We plan to extract two types of data from the CDR, the first pertaining to a single tower* (*measuring the number of incoming/outgoing calls*), *and the second concerning a pair of two towers* (*measure the flow of calls between them*). The raw CDR data contains each cell tower's location information in latitude and longitudes. We intend to work at the spatial granularity of the *Voronoi areas* associated with cell tower placements. To successfully use such data, telecommunication providers just need to share anonymized, aggregated call detail records in a regulated manner. Early hints of this are already visible (e.g., the D4D Challenge[3]), and different frameworks are being developed to attract even more providers to join the endeavour [Parate and Miklau, 2009].

Other Data Sources as Indicators of Poverty

Data from remote sensing and communication networks form the primary proxy data source for poverty estimation. Other variables that may contribute towards poverty are the environmental, political, and cultural characteristics of a region. For instance, natural calamities, such as floods, droughts, tornadoes, extreme/high low rainfall, and famines could result in poverty aggravation. Furthermore, the details about the ruling party of a region, the policies it has implemented, and quantifying its progress could also shed light on the region's economic development. Population-related statistics, such as the number of women, men, children, and senior citizens in an area might as well have some bearing on its economic condition. Information from local news snippets could also shed light on regional developments, crime rates, and other advancements. It also remains an open question whether cultural information, such as characteristics of different tribes, the different caste and their distribution, plays an active part in poverty tracking, assessment, and expansion. Street-view images of an area over time could also help depict its developments or degradation. *We intend to consider these ancillary factors and determine whether any of these positively correlate with poverty deprivation.*

Data Integration

The distinctive aspect of this study is that, rather than relying solely on new proxy data sources for poverty estimation, we integrate them with traditional census and household survey data. The survey datasets provide vital individual-level information about age, sex, religion, caste, mother tongue, marital status, education, disability status, land ownership, irrigation infrastructure, and tenancy status, to name a few. They also include details on the availability of bathrooms, drinking water, separate kitchen, electricity, electronic devices, vehicle count, cooking fuel, and roof, wall, and floor materials, among other things[4]. This results in a diverse dataset containing images, text, numeric data, and categorical labels. While traditional statistical methods cannot accommodate non-numeric data inputs, the current state-of-the-art deep learning architectures are perfectly suited for processing and merging unstructured data from disparate sources. We elaborate on this in the following section.

3.2 Learning and Reasoning Poverty Estimation

Neural Models for Poverty Estimation

Deep neural models are being aggressively employed to attain various sustainable development goals [Vinuesa *et al.*, 2020; Vinuesa and Sirmacek, 2021]. For poverty assessment and tracking as well, there have been numerous successful attempts of using ML methods utilizing traditional [Isnin@Hamdan *et al.*, 2020; niu, 2020] as well as proxy data sources [Jean *et al.*, 2016; Smith-Clarke *et al.*, 2014]. However, none of these methods designs systems that aggregate and combine different data sources. To address this, *we aim to present a multi-input neural architecture that can aggregate and combine information from several disparate data sources in a non-trivial fashion*. Figure 2 presents a schematic diagram of our generalized deep neural model that takes various inputs, such as images, text, as well as numeric and categorical data. The numeric data derived from the census and household surveys can be easily fed to a deep neural model by combining all the features into a vector representation. Categorical attributes, such as gender, education, religion, caste, and marital status can be encoded using a technique, called *entity embeddings* [Guo and Berkhahn, 2016]. Unstructured data types of images and text snippets cannot be fed directly into neural models. While deep learning models specialize in processing unstructured data, they require input in the form of numeric representations. However, as we have limited training data, training visual or textual embeddings from scratch will, in all likelihood, not yield desirable results. Thus, *we propose to use a transfer learning approach wherein we use pre-trained models trained on large corpora and fine-tune them to our task-specific dataset*. For the visual inputs of satellite images, we can utilize pre-trained CNN-based models, such as VGG Net [Simonyan and Zisserman, 2014], or ResNet [He *et al.*, 2016]. For the textual inputs, we prefer the language models, such as BERT [Devlin *et al.*, 2019] or RoBERTa [Liu *et al.*, 2019]. Each of the representations goes through a non-linear transformation before the data merging operation. The data merging can be done in various ways, such as early fusion, late fusion, naive concatenation, attention mechanism [Bahdanau *et al.*, 2014], or gating-based fusion. We treat our problem as a classification task wherein we have four ground-truth labels at the district level – 'advanced', 'catching up', 'falling behind', and 'lagged'. This is calculated based on aggregating the individual MDPI scores at the district level.

Correlation, Association, and Causality

Analysing the relationship of the input variables with the target output can be a challenging step but it is important for strategic actions. Such insights are important to determine the driving factors (causative), factors that exhibit linear relationships (correlative), and factors that co-occur (associa-

[3] http://www.d4d.orange.com/home

[4] https://censusindia.gov.in/census_and_you/data_item_collected_in_census.aspx

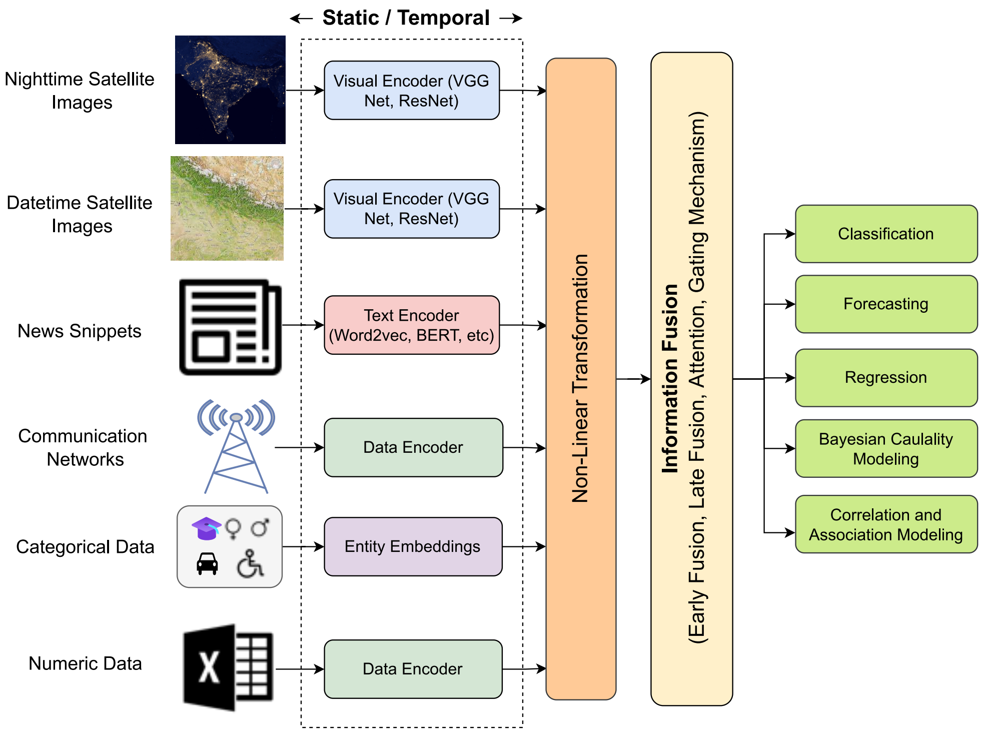
← Static / Temporal →
Nighttime Satellite Images
Datetime Satellite Images
News Snippets
Communication Networks
Categorical Data
Numeric Data
Visual Encoder (VGG Net, ResNet)
Visual Encoder (VGG Net, ResNet)
Text Encoder (Word2vec, BERT, etc)
Data Encoder
Entity Embeddings
Data Encoder
Non-Linear Transformation
Information Fusion
(Early Fusion, Late Fusion, Attention, Gating Mechanism)
Classification
Forecasting
Regression
Bayesian Caulality Modeling
Correlation and Association Modeling

Figure 2: Proposed multi-input deep learning model for aggregating and processing data from proxy and traditional data sources.

tive). We propose to use the Bayesian Networks for identifying causative factors, Pearson's correlation to determine the correlative factors, and the Hypergeometric test to discover the associative factors. Such correlative, associative, and causal analyses will aid in shaping and designing efficient policymaking.

3.3 Temporal Analysis for Poverty Estimation

While the current vogue is to predict poverty using static data properties, this approach does not encapsulate the concept of poverty in its entirety. Poverty is a complex phenomenon with a very strong temporal component. Furthermore, recent studies have demonstrated that ML models trained for poverty prediction have poor time transferability, i.e., whether models built on data from one year can make sound predictions on data from another year [Bansal *et al.*, 2020]. Thus, for efficient and robust poverty assessment, we must consider and process data in a temporal fashion. For example, a consistent decline in rainfall in an area over time might actively contribute toward increased chronic poverty. However, outlier events, such as natural calamities or socio-political upheavals are rare phenomena leading to only ephemeral impoverishment and transient poverty as people finally recover. Thus, in this project, *we propose a new research direction and methodology for the temporal assessment and prediction of poverty using longitudinal dissection of data*. To start with, we plan to consider a time frame of 5 years for the temporal analysis. State-of-the-art sequences learning models, such as LSTMs and GRUs have been very successful in capturing temporal dependencies and sequential data. The range of transformer-based architectures [Vaswani *et al.*, 2017] succeed likewise in this using the self-attention mechanism. Using these architectures, we will train a neural network using *teacher-forcing* in which, during training, the model receives the ground-truth output $y(t)$ as input at time $t + 1$. Thus, for each time step, the model receives the standard inputs, along with the output from the previous time step.

4 Model Evaluation

When evaluating a deep learning model for multifaceted and longitudinal data for poverty estimates and livelihood capabilities, it is important to consider several criteria. For our case we focus on the following six criteria:

1. **Accuracy:** The model should be able to accurately predict poverty estimates and livelihood capabilities based on the given data.
2. **Bias:** The model should not be biased towards any particular group or demographic. The model should be fair and impartial in its predictions.
3. **Generalizability:** The model should be able to generalize well to new data that it has not seen before. This is important because poverty estimates and livelihood capabilities can vary across different regions and populations. Thus, the model needs to be able to capture these variations.
4. **Interpretability:** The model should be interpretable, meaning that it should be possible to understand how the model arrived at its predictions. This is important for ensuring that the model is not making predictions based on

irrelevant or biased factors. Moreover, to design effective policies, the model interpretability can help provide causal relations between the input data features and the poverty prediction outcome.

5. **Robustness:** The model should be robust to changes in the data and any noise or errors that may be present. This is important because poverty estimates and livelihood capabilities can be affected by various factors, and the model should be able to handle these variations.
6. **Scalability:** The model should be scalable, meaning that it should be able to handle large volumes of data efficiently. This is important because poverty estimates and livelihood capabilities data can be vast and complex, and the model should be able to handle this complexity.

5 Challenges

Though poverty amelioration has been a primary priority of the Indian government, the fact of the matter remains that millions of Indians continue to be poor by national and international standards, despite persistent efforts since independence. The utilization of AI/ML techniques may introduce additional challenges for efforts regarding poverty assessment. They are elaborated as follows.

1. **Lack of labeled data:** Deep learning models require large amounts of labeled data to achieve accurate predictions. The more data you feed them, the better results they are likely to yield. However, in many cases, poverty-related data is not readily available or is difficult to label accurately. Therefore, this requires substantial efforts in data gathering, storing, and processing.
2. **Data quality:** Poverty-related data can be subject to quality issues such as missing values, errors, and inconsistencies, leading to biased results. or example, there may be missing data on income or consumption, or the data may not be representative of the population being studied.
3. **Data bias:** Data bias can occur when the training data used to train deep learning models is not representative of the population being studied. This can lead to inaccurate predictions and exacerbate existing inequalities.
4. **Data privacy:** Poverty-related data may contain sensitive information about individuals or households, making it difficult to collect and share. This can limit the availability of data for training deep learning models.
5. **High Carbon Footprint:** the AI algorithms processing big data have high energy requirements and carbon footprints, which can have a detrimental impact on SDG 7 (Affordable and Clean Energy) and SDG 13 (Climate Action).
6. **Lack of skilled personnel:** Finally, designing and monitoring AI-based systems necessitates experts in these fields. Moreover, as AI/ML is a rapidly evolving discipline, the recruited employees must regularly update their understanding of this field. However, India, which otherwise produces thousands of IT workers each year, has a severe scarcity of AI skills professionals.

6 Risks, Limitations, Ethical Considerations

Poverty is a structural phenomenon, and tracing its causal factors over a time period might not be enough to articulate the causes of deprivation. The macro statistics, including budget allocation by government, public and private investment, and revenue generation of respective states, are not included in the analysis. Qualitative information like life history, challenges, and opportunities at the household level are not incorporated into the analysis. The data integration might reduce the quality and explainability of the individual datasets. The data used in the study are accessible from public platforms. For satellite images, outputs will be published at the district level.

7 Expected Results and Long-Term Plans

The holistic results of this study will provide a cogent understanding of the underlying causes of poverty and the capability of lagged regions to come out of it. Socio-economic factors in terms of access to amenities, education, health, connectivity, and living conditions facilitate livelihood capabilities in rural India. Inability to generate sustainable livelihood results in pushing the lagged region behind. The proxy indicators have the potential for a deeper understanding of poverty. The traditional poverty measurement indicators of income level or consumption expenditure also have the probability of reporting error, whereas the analysis of poverty from different facets will assist policymakers in eradicating poverty. Even the advanced states have intrastate differences, and the districts with a concentration of marginalized populations are the most lagged regions of the state. The integrated results can provide a different perspective of poverty based on the quality of life and livelihood capabilities and strengthen the outcome of conventional indicators. Policy suggestions will be based on the causal relationship of socioeconomic determinants of quality of life based on longitudinal data. In this context, the current social welfare and rural development schemes will be evaluated based on the results.

8 Conclusion

Regional poverty has been a major concern since the Independence period in India. The polarised growth in advanced states does not have the ripple effect of growth for low-income states. The proposal aims to understand the social, economic, regional, and cultural dimensions of uneven development and its impact on the poverty condition of the lagged region. The application of AI in targeting lagged regions has immense scope to identify the explaining factors of poverty. The analyses also trace the quality of life and development indicators of the advanced region and the catching-up region, which will suggest a path of development for the lagged region. Data integration is a powerful method to measure poverty rather than using only traditional or proxy variables. Poverty and inequality would be widening in developing countries due to an increase in population and growth-centred policies. Therefore, targeting the lagged region and the vulnerable population is essential to eradicate poverty and improve the quality of life to achieve the goal of "zero poverty".

References

[Addison and Stewart, 2015] Douglas M Addison and Benjamin Stewart. Nighttime lights revisited: the use of nighttime lights data as a proxy for economic variables. *World Bank Policy Research Working Paper*, (7496), 2015.

[Ahluwalia *et al.*, 1980] Montek S. Ahluwalia, John H. Duloy, Graham Pyatt, and T. N. Srinivasan. Who benefits from economic development? comment. *The American Economic Review*, 70(1):242–245, 1980.

[Alkire and Foster, 2011] Sabina Alkire and James Foster. Counting and multidimensional poverty measurement. *Journal of Public Economics*, 95(7):476–487, 2011.

[A.M, 2020] Sharath A.M. The challenges of poverty, types and its causes. *International Research Journal on Advanced Science Hub*, 2:81–85, 2020.

[Atkinson, 2003] Anthony B Atkinson. Multidimensional deprivation: contrasting social welfare and counting approaches. *The Journal of Economic Inequality*, 1(1):51–65, 2003.

[Ayush *et al.*, 2020] Kumar Ayush, Burak Uzkent, Marshall Burke, David Lobell, and Stefano Ermon. Generating interpretable poverty maps using object detection in satellite images. In *IJCAI*, pages 4410–4416, 7 2020.

[Ayush *et al.*, 2021] Kumar Ayush, Burak Uzkent, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Efficient poverty mapping from high resolution remote sensing images. *AAAI*, 35:12–20, May 2021.

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Balaji, 2020] M Balaji. Negotiating poverty line-study on density effect around the poverty line for indian states. *The Singapore Economic Review*, 65:139–160, 2020.

[Bank, 2020] World Bank. *Poverty and shared prosperity 2020: Reversals of fortune*. 2020.

[Bansal *et al.*, 2020] Chahat Bansal, Arpit Jain, Phaneesh Barwaria, Anuj Choudhary, Anupam Singh, Ayush Gupta, and Aaditeshwar Seth. Temporal prediction of socio-economic indicators using satellite imagery. In *CoDS COMAD*, page 73–81, 2020.

[Blumenstock *et al.*, 2015] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

[Chaudhuri *et al.*, 2013] Basudeb Chaudhuri, Namrata Gulati, Apara Banerjee, Ahana Roy, Imdadul Halder, Safayet Karim, and Paul Vertier. Multi-dimensional poverty index-a state level analysis of india. 2013.

[Das *et al.*, 2021] Pinaki Das, Sudeshna Ghosh, and Bibek Paria. Multidimensional poverty in india: a study on regional disparities. *GeoJournal*, pages 1–20, 2021.

[Dehury and Mohanty, 2015] Bidyadhar Dehury and Sanjay K Mohanty. Regional estimates of multidimensional poverty in india. *Economics*, 9(1), 2015.

[Dev, 2010] S. Mahendra Dev. *Inclusive Growth in India: Agriculture, poverty and human development*. 2010.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, June 2019.

[Ghosh *et al.*, 2013] Tilottama Ghosh, Sharolyn J Anderson, Christopher D Elvidge, and Paul C Sutton. Using nighttime satellite imagery as a proxy measure of human wellbeing. *sustainability*, 5(12):4988–5019, 2013.

[Guo and Berkhahn, 2016] Cheng Guo and Felix Berkhahn. Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[Hofer *et al.*, 2020] Martin Hofer, Tomas Sako, Arturo Martinez Jr, Mildred Addawe, Joseph Bulan, Ron Lester Durante, and Marymell Martillan. Applying artificial intelligence on satellite imagery to compile granular poverty statistics. *Asian Development Bank Economics Working Paper Series*, (629), 2020.

[Isnin@Hamdan *et al.*, 2020] Rusnita Isnin@Hamdan, Azuraliza Abu Bakar, and Nur Samsiah Sani. Does artificial intelligence prevail in poverty measurement? *Journal of Physics: Conference Series*, 1529(4):042082, 2020.

[Jean *et al.*, 2016] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Mishra and Ray, 2013] Ankita Mishra and Ranjan Ray. Multi-dimensional deprivation in india during and after the reforms: Do the household expenditure and the family health surveys present consistent evidence? *Social Indicators Research*, 110(2):791–818, 2013.

[Mohanty, 2011] Sanjay K Mohanty. Multidimensional poverty and child survival in india. *Plos one*, 6(10):e26857, 2011.

[niu, 2020] Measuring urban poverty using multi-source data and a random forest algorithm: A case study in guangzhou. *Sustainable Cities and Society*, 54:102014, 2020.

[Nussbaum and Sen, 1993] Martha Nussbaum and Amartya Sen. *The quality of life*. Clarendon Press, 1993.

[Parate and Miklau, 2009] Abhinav Parate and Gerome Miklau. A framework for safely publishing communication traces. In *CIKM*, page 1469–1472, 2009.

[Pokhriyal *et al.*, 2015] Neeti Pokhriyal, Wen Dong, and Venu Govindaraju. Virtual networks and poverty analysis in senegal. *arXiv preprint arXiv:1506.03401*, 2015.

[Puttanapong *et al.*, 2020] Nattapong Puttanapong, Arturo Martinez Jr, Mildred Addawe, Joseph Bulan, Ron Lester Durante, and Marymell Martillan. Predicting poverty using geospatial data in thailand. *Asian Development Bank Economics Working Paper Series*, (630), 2020.

[Sarkar, 2012] Sandip Sarkar. Multi-dimensional poverty in india: Insights from nsso data. *Resource document*, 14:2017, 2012.

[Sheehan *et al.*, 2019] Evan Sheehan, Chenlin Meng, Matthew Tan, Burak Uzkent, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Predicting economic development using geolocated wikipedia articles. In *SIGKDD*, page 2698–2706, 2019.

[Shi *et al.*, 2020] Kaifang Shi, Zhijian Chang, Zuoqi Chen, Jianping Wu, and Bailang Yu. Identifying and evaluating poverty using multisource remote sensing and point of interest (poi) data: A case study of chongqing, china. *Journal of Cleaner Production*, 255:120245, 2020.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Škare and Družeta, 2016] Marinko Škare and Romina Pržiklas Družeta. Poverty and economic growth: a review. *Technological and Economic development of Economy*, 22(1):156–175, 2016.

[Smith-Clarke *et al.*, 2014] Christopher Smith-Clarke, Afra Mashhadi, and Licia Capra. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 511–520, New York, NY, USA, 2014. Association for Computing Machinery.

[Smith-Clarke, 2021] Christopher Smith-Clarke. *Estimating poverty maps from aggregated mobile communication networks*. PhD thesis, UCL (University College London), 2021.

[Soto *et al.*, 2011] Victor Soto, Vanessa Frias-Martinez, Jesus Virseda, and Enrique Frias-Martinez. Prediction of socioeconomic levels using cell phone records. In *UMAP*, pages 377–388, 2011.

[Steele *et al.*, 2017] Jessica E Steele, Pål Roe Sundsøy, Carla Pezzulo, Victor A Alegana, Tomas J Bird, Joshua Blumenstock, Johannes Bjelland, Kenth Engø-Monsen, Yves-Alexandre De Montjoye, Asif M Iqbal, et al. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127):20160690, 2017.

[Tripathi and Yenneti, 2020] Sabyasachi Tripathi and Komali Yenneti. Measurement of multidimensional poverty in india: A state-level analysis. *Indian Journal of Human Development*, 14(2):257–274, 2020.

[Van der Berg *et al.*, 2006] Servaas Van der Berg, Ronelle Burger, Rulof Burger, Megan Louw, and Derek Yu. Trends in poverty and inequality since the political transition. 2006.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, volume 30, 2017.

[Vinuesa and Sirmacek, 2021] Ricardo Vinuesa and Beril Sirmacek. Interpretable deep-learning models to help achieve the sustainable development goals. *Nature Machine Intelligence*, 3(11):926–926, 2021.

[Vinuesa *et al.*, 2020] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. The role of artificial intelligence in achieving the sustainable development goals. *Nature communications*, 11(1):1–10, 2020.

[Xie *et al.*, 2016] Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Transfer learning from deep features for remote sensing and poverty mapping. In *AAAI*, page 3929–3935, 2016.