Poverty Mapping in India using Machine Learning and Deep Learning Techniques

Srishti Gulecha R[1], Muthu Reshmi K[1], Rishitha N[1], Vani K[2]

[1] Dept. of Information Science and Technology, College of Engineering Guindy, Chennai, India - (muthureshmi08, nrishitha58, srishtigulecha02)@gmail.com

[2] Dept. of Information Science and Technology, College of Engineering Guindy, Chennai, India - vani@annauniv.edu

Keywords: Poverty Estimation, Geospatial Big Data, Deep Learning, Machine Learning, Image Processing, OpenStreetMap

Abstract

Poverty remains a persistent global challenge, that affects millions worldwide and hinders sustainable development goals. Poverty related data is traditionally collected by an on-the-ground household survey which is conducted once in a few years. Unfortunately, in India, the reach of this conventional data collection method has many limitations, and it is costly, time-consuming, and laborious. The study will assist in identifying areas of poverty and also the levels of poverty which will help policymakers in creating policies that will improve such areas. The research leverages a rich variety of data sources which includes satellite imagery, geospatial data, socio-economic surveys and Point of Interest (POI) data. To extract meaningful patterns and correlations within this diverse dataset, various machine learning and deep learning algorithms such as Decision Tree Regressor, Random Forest Regressor, Convolutional Neural Networks (CNNs) and Multi Layer Perceptron (MLP) are employed. With the help of Random Forest Regressor, the study was able to estimate the poverty at village/town level with a R2-score of 0.778.

1. Introduction

Poverty is a complex and persistent issue in India. For governments and humanitarian groups to monitor the advancement of bettering livelihoods, accurate local-level poverty measuring is a crucial responsibility. Poverty mapping is a crucial tool that helps us understand where poverty is most severe and why. By using geospatial data (Avtar et al., 2019) and deep learning techniques (Aprianto et al., 2022, Babenko et al., 2017, Jean et al., 2016, Luo et al., 2022), we can identify areas in need and create targeted solutions to reduce poverty and improve the lives of those who are most vulnerable.

Poverty mapping relies on various data sources, including household surveys and census data. Data on income, education, healthcare access, and housing conditions are collected to assess the poverty levels in different regions. Satellite data can be used for analysis, allowing researchers to track changes over time and understand trends in poverty dynamics. Machine learning algorithms can efficiently process and analyze vast amounts of data, enabling the identification of complex patterns and relationships. GIS allows for the creation of informative maps and visualizations, making complex data more accessible and understandable. Using these methods is often more cost-effective than conducting extensive ground surveys.

The primary objective of this study is to systematically identify and map areas affected by poverty using a multi-faceted approach. To achieve this, a wide range of data sources and technologies are harnessed. Satellite imagery (Puttanapong et al., 2022, Putri et al., 2023, Tingzon et al., 2019) can be used to extract factors like vegetation index (NDVI), water availability (NDWI), built up area (NBI), pollution levels (CO, SO2, NO2), nighttime light intensity, and land surface temperature. Additionally, data can be extracted from OpenStreetMap (Lin Htet et al., 2021, Puttanapong et al., 2022, Putri et al., 2023, Hu et al., 2022) to identify key points of interest (POI) like schools, hospitals, and markets, which are integral to assessing poverty. Building footprint data will be used to estimate the number of buildings which can also be used to predict poverty (Luo et al., 2022, Ayush et al., 2020, Tingzon et al., 2019). Furthermore, machine learning and deep learning models will be applied to predict poverty levels at the district and village/town level based on the collected data. Finally, visual maps are created that illustrate poverty levels in different regions, providing a valuable resource for policymakers and organizations to make informed decisions for poverty alleviation.

2. Related Works

This section highlights related works which use satellite images and geospatial data for extracting various features for poverty prediction with the help of machine learning and deep learning techniques. It gives an overview about the challenges involved in developing this overall study.

High-resolution daytime satellite imagery was used to predict local-level consumption expenditure using object detection to generate poverty maps in Uganda (Ayush et al., 2020). In the study, YOLOv3 model is trained using xView Dataset to detect objects like building, passenger vehicle, truck, railway vehicle, construction site, etc. from satellite images. Regression models like Gradient Boosting, Decision Trees and Linear Regression are trained with cluster level extracted counts of detected objects as features for prediction of consumption expenditure which is used to compute poverty statistics.

Various factors like the intensity of nighttime light(NTL), land cover, vegetation index, land surface temperature, built-up areas, and points of interest are considered to predict poverty in Thailand (Puttanapong et al., 2022). The study states that NTL acts as a proxy measure of level of economic activity in a given area. Data was obtained from Google Earth Engine, open cloud-based data storage and computing platform which provides access to satellite imagery for free. Point Of Interest (POI) data was obtained from OpenStreetMap. Among the 4 techniques used for prediction, Random Forest Regressor produced the best results.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-2024-319-2024

The work (Putri et al., 2023) proposes a novel approach to generate poverty maps for East Java, Indonesia using machine learning and deep learning approaches. The first approach involves the use of multisource satellite images and point of data utilising zonal statistics feature extraction. Indicators of poverty derived from images include Normalised Difference Water Index (NDWI), Normalised Difference Vegetation Index(NDVI), Built-Up Index (BUI), Carbon nonoxide (CO), Nitrogen dioxide(NO2), and Sulphur dioxide (SO2) for detecting economic activity, LandSurface Temperature (LST), and Nighttime Light (NTL) intensity. Google Earth Engine and OpenStreetMap were used to collect and process the data. In the second approach, the deep learning architecture of Resnet-34 transfer learning feature extraction is used to build the model from daytime multiband and nighttime light intensity satellite images. The CNN-1D model was determined to be the best model in the first scenario and the Resnet-34 + MLP model to be the best model in the second scenario.

The research (Tingzon et al., 2019) implemented an approach using geospatial data including nighttime lights imagery, daytime satellite imagery, OpenStreetMap data and human settlement data for estimating socioeconomic indicators for Philippines. They fine-tune a VGG16-based model to classify nighttime light intensity from satellite images. A ridge regression model is used to map cluster-level features to socioeconomic indicators. A fusion of OSM-Nightlights hybrid model achieved better results.

The study (Hu et al., 2022) provides an approach to village-level poverty identification using satellite imagery and geospatial data. Their approach integrates high-resolution imagery (HRI), Point Of Interest (POI), OpenStreetMap (OSM), and digital surface model (DSM) data. HRI images are used for land use classification. The time cost to the nearest facilities and services combined with HRI, POI, and OSM is calculated. And the the dispersity of village settlements is also used. Random forest algorithm was used to model the relationship between these variables and predict the poverty incidence as poor, medium or wealthy.

The work (Babenko et al., 2017) estimates poverty directly from high and medium resolution satellite images for Mexico. The study uses satellite imagery provided by both Planet and Digital Globe. Out of two CNN based architectures that were used, GoogleNet architecture outperformed the VGG architecture.

A method that combines deep learning along with satellite imagery for poverty prediction was also proposed (Jean et al., 2016). The study maps daytime satellite photos to equivalent nighttime light intensity levels using convolutional neural networks. Poverty mapping estimation was done by applying transfer learning Resnet-34 feature extraction on daytime satellite data for five African countries - Nigeria, Tanzania, Uganda, Malawi and Rwanda.

3. Methodology

The primary objective of this study is to systematically identify and map areas affected by poverty using a multi-faceted approach. The study is composed of 4 different modules - data collection and processing, zonal level feature extraction, model training and poverty mapping as shown in Figure 1.

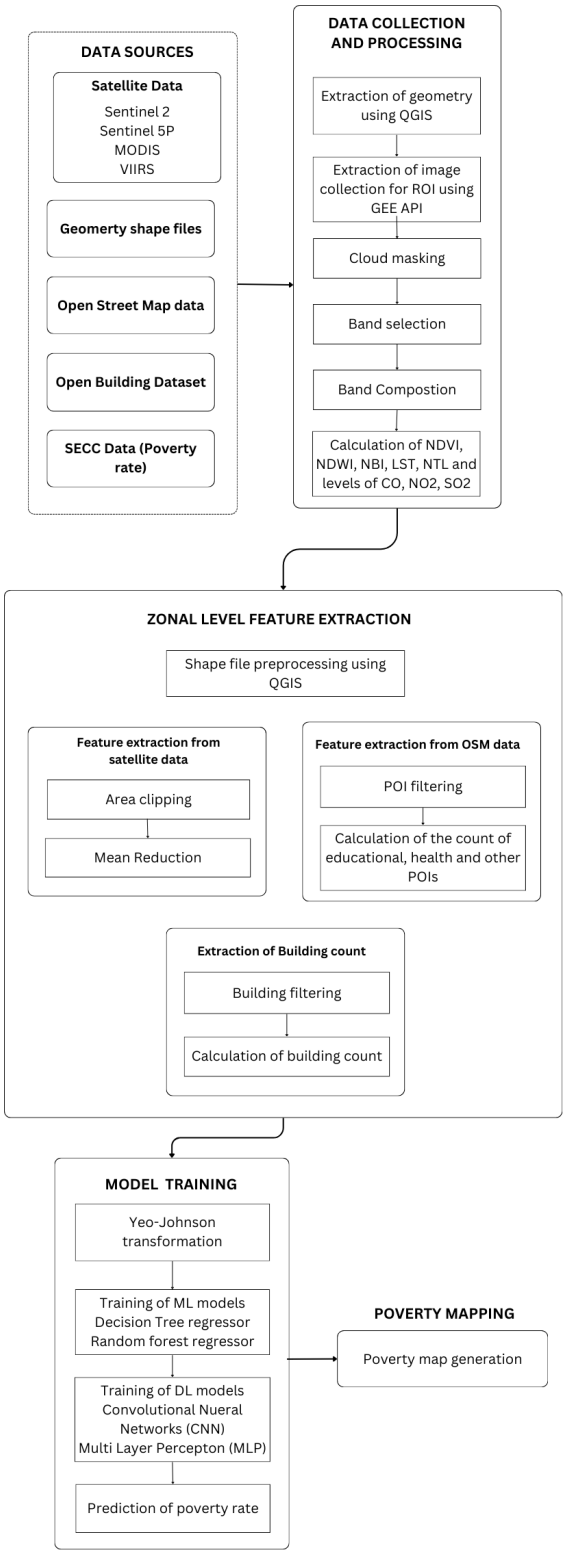


Figure 1. System Architecture

3.1 Data Collection And Processing

Figure 2 describes the process for data collection and processing. The Google Earth Engine API is used to access the images from various satellites. These images are then processed to mask out the clouds. For performing cloud masking, the band with cloud information is chosen and the bits with clouds are masked out. After performing cloud masking on the required

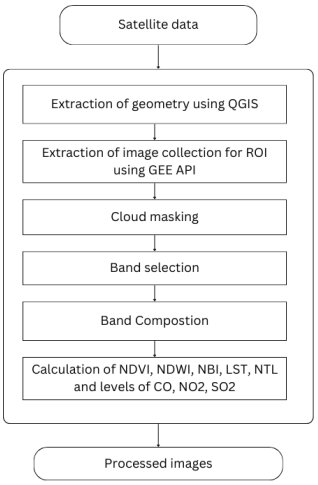


Figure 2. Data Collection And Processing

images, the proper bands are selected and composed to get the required features.

The bands used from each of the satellite sources to obtain the respective features are mentioned in the Figure 3.

Image collection	Resolution	Features	Bands
Sentinel 2	**60 m** spatial resolution	NDVI NDBI NBI	Band 4 (red), Band 8 (NIR) Band 3 (Green), Band 8 (NIR) Band 4 (red), Band 8 (NIR), Band 11 (SWIR)
Sentinel 5P	**1113.2 meters** spatial resolution	CO NO2 SO2	CO_column_number_density NO2_column_number_density SO2_column_number_density
MODIS	**1,000 m** spatial resolution	LST	LST_Day_1km
VIIRS	**750 m** spatial resolution	NTL	avg_rad

Figure 3. Band Selection

The bands used for NDVI are bands 4(NIR) and 8(Red), for NDWI bands 4(NIR) and 3(Green) are used and for NBI bands 8(NIR) and 11(SWIR) are used. While for LST and NTL, the following bands are choosen - LST Day 1km and avg rad respectively. The bands CO column number density, NO2 column number density, SO2 column number density are used for getting the CO, NO2 and SO2 respectively. Band Composition is done for NDVI using the equation 1. Equation 2 is used for calculation of NDWI. NBI is calculated using the equation 3.

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{1}$$

$$NDWI = \frac{GREEN - NIR}{GREEN + NIR} \tag{2}$$

$$NBI = \frac{SWIR * RED}{NIR} \tag{3}$$

Cloud masking is needed for Sentinel-2 and MODIS images. For Sentinel-2,the QA band (’QA60’) contains pixel quality information. Control Band (QC Day Band) from the input image, containing information on pixel quality is used for cloud masking in MODIS. After taking the corresponding bands for the respective satellite images,binary masks are defined for clouds and cirrus using bitwise left shifts. By combining these masks using the bitwiseAnd method, it identifies clear pixels with values of 0 in both masks. This binary mask is applied to the input image to effectively mask out cloudy and cirrus-affected pixels.

For the district level, the state Tamil Nadu is chosen for the study. There are a total of 32 districts according to the 2011 census. Data Processing is performed for each of these districts. The processed images are then saved as tiff file.

For mapping poverty at the village/town level, the following districts-Kanchipuram, Nawada, Namakkal, Gadag, Thiruvarur, Alrajpur, Bahraich, Srikakulam, Balrampur, Salem, and Madurai are chosen.

3.2 Zonal Level Feature Extraction

After data collection and processing is done, the processed images are available to perform feature extraction at two different zonal levels - district and village/town as shown in Figure 4. For clipping the image to the required region, the shape file for that region is used. The shape files are taken from the The Socioeconomic High-resolution Rural-Urban Geographic Platform for India (SHRUG) (Asher et al., 2021). The shape file for specific regions is extracted with the help of QGIS software. The features extracted from satellite images are NDVI, NDWI, NBI, CO, NO2, SO2, NTL and LST.

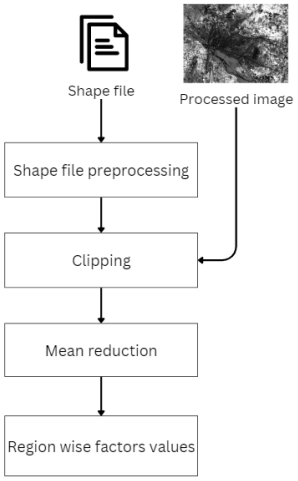


Figure 4. Zonal Level Feature Extraction

For feature extraction of each of the districts of Tamil Nadu, the saved TIFF files are loaded into QGIS. For each feature, the corresponding TIFF files of all districts are selected and merged. This is done for each of the features. The mean for the features is obtained by using the zonal statistics tool available in QGIS. The values are then exported and saved as a .xlsx file.

At village/town level feature extraction process for satellite data we compute feature values for each region within a district. For every region, the image is clipped to the specified area, and the mean value for that region is calculated. Subsequently, the calculated values for each region are stored in a dataframe. Finally, the dataframe is saved as a .xlsx file, producing an output file with the computed feature values for each region.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-2024-319-2024 | © Author(s) 2024. CC BY 4.0 License.

Another set of important features, i.e the POI, which comprises a huge number of important locations, can reflect the density of economic activity and accessibility of a region. Features extracted from this are: Count of educational POIs, Count of healthcare POIs, Count of other POIs like banks, restaurants, hotels, theatres, etc. The python osmnx module is used for accessing OpenStreetMap(OSM) data. For district level, the total count of POIs for each category within every district is calculated. Meanwhile for village level, a centroid is taken and a fixed buffer is created around it, representing the region. Within the buffer zone of 2.5 kms, the OSM data points are filtered according to the selected categories and the count of POIs are calculated for each category within the defined region. The result is a dataset that provides region-specific information on the number of health, education, and other POIs.

The feature Building count of a region is computed using the Google Earth Engine Open Building dataset. After collecting the dataset the region (AOI) is defined to get the building count of it and the images are filtered based on this AOI. The count is taken by choosing an confidence level (greater than 0.5) that selects the buildings from the satellite images. To get the building count at the district level, the sum of building counts at the village/town level of that district is taken.

3.3 Model Training

The features extracted previously - NDVI, NDWI, NBI, CO, NO2, SO2, LST, NTL, various POIs and building count are the independent variables. The poverty rate obtained from SECC(Ministry of Rural Development, 2011) is the dependent variable. The Figure 5 describes this module.

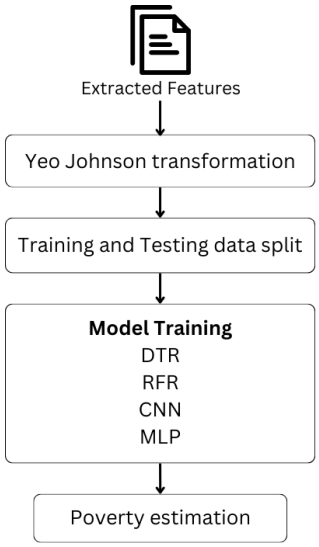


Figure 5. Model Training

Before training a model, the data is cleaned by removing null values. After which it is transformed, and for this purpose Yeo-Johnson transformation is used. The reason for using this transformation is that it inflates low variance data and deflates high variance data to create a more uniform dataset and can handle negative values as well. This transformed data is then used for model training. The data is split into training and testing dataset. Decision Tree Regressor(DTR), Random Forest Regressor(RFR), CNN and MLP are trained and tested using training data and testing data respectively.

The Convolutional Neural Network(CNN) architecture has been developed using Keras. The layers used in showw in Figure 6. The initial layer, constitutes a 1D convolutional layer with 128 filters, a kernel size of 3, ReLU activation, and L2 regularization applied to the kernels. Subsequently, a max-pooling layer follows with the window of size 2. An additional convolutional layer with 64 filters, a kernel size of 3, ReLU activation, and L2 regularization. Another max-pooling layer further reduces dimensionality before the data is flattened to be fed into a dense layer. Two additional dense layers are included, each comprising 128 and 64 units respectively, with ReLU activation and L2 regularization. The output layer uses L2 regularization.

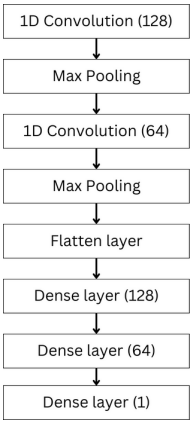


Figure 6. Layers used in CNN Model

The Multilayer Perceptron(MLP) model was constructed using a sequential architecture in Keras. The layers used is shown in Figure 7. The first layer, comprises a densely connected layer with 64 units and rectified linear unit (ReLU) activation function. The subsequent layer, is another densely connected layer with 32 units and ReLU activation. Lastly, the output layer, consists of a single unit. Once the models are trained, they are used to make predictions regarding the poverty rate for each region.

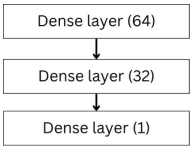


Figure 7. Layers used in MLP Model

To analyze the performance, RMSE - Root Mean Squared Error, MAE - Mean Absolute Error and R2 score are then calculated for all the models and the one with less error rates will be chosen as the best model.

3.4 Poverty Mapping

A poverty map is created for Tamil Nadu using QGIS. The trained model is used to predict the district level poverty rates which is used for visualization. These predicted rates are then categorized into intervals of 5, and each interval is assigned a color. Two layers are added - predicted poverty rate and actual poverty rate along with a legend. Labelling is done for each region with district’s name and poverty rate for quick inference.

The creation of poverty maps for village/town level begins by utilizing the best model trained among the random forest regressor, Decision Tree regressor, CNN and MLP to predict the

poverty rate for the input region. These predicted rates are then categorized into intervals of 10, each interval has a specific colour thus effectively classifying them. The final step involves the use of QGIS. Poverty map is generated for all the districts mentioned earlier at village/town level. There are two layers added one with the actual poverty rate and the other with the predicted poverty rate along with legend and appropriate labels.

4. Results and Discussions

Initially, data collection is done from multiple satellites using Google Earth Engine API. Cloud masking is performed for images derived from Sentinel-2 and MODIS and a sample is shown in Figure 8.

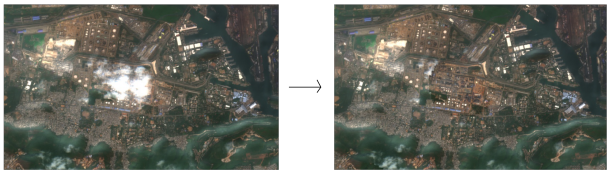


Figure 8. Before and after cloud masking in Sentinel 2 images

Figure 9. Cloud masking for Kanchipuram Region

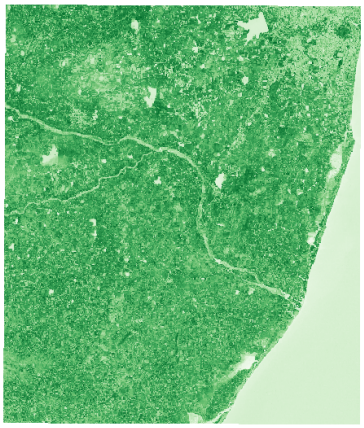


Figure 10. NDVI for Kanchipuram

After selecting appropriate bands from images, various indices like NDVI, NDWI and NBI and other factors are calculated from satellite images as mentioned previously. Figure 9 shows the cloud masked image for Kanchipuram which is used to calculate indices. Figure 10, Figure 11 and Figure 12 are the visual representations of the calulated NDVI, NDWI and NBI respectively. Similarly the features - CO, NO2, SO2, NTL and LST are also calculated for all the 11 districts.

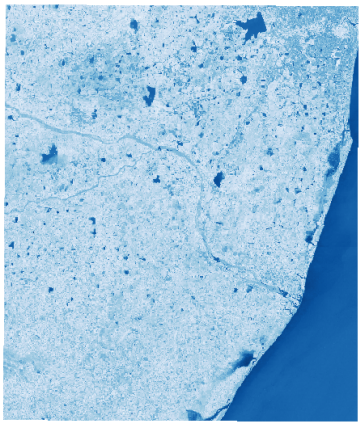


Figure 11. NDWI for Kanchipuram

Figure 12. NBI for Kanchipuram

Figure 13 and Figure 14 visualizes the calculated village level values for two features namely NTL intensity and NDVI respectively for villages/towns of Kanchipuram. Similarly, other features are also calculated for each district’s villages/towns.

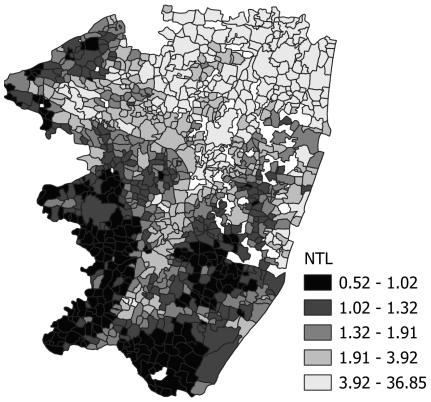


Figure 13. NTL for Kanchipuram - Visualization at Zonal Level

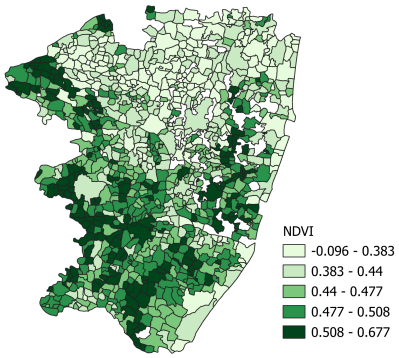


Figure 14. NDVI for Kanchipuram - Visualization at Zonal Level

POI data is extracted from OSM and building count is evaluated using GEE Open building dataset for each region. For Kanchipuram region, the Figure 15 shows the various POIs.

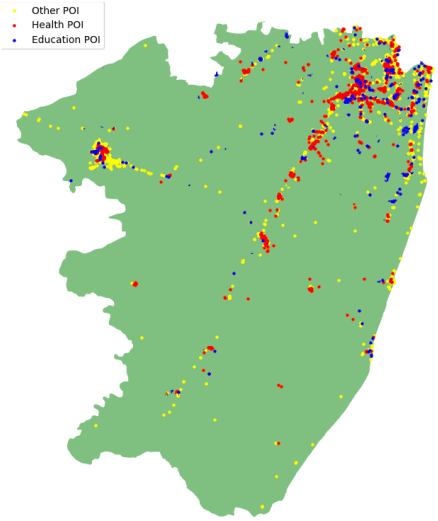


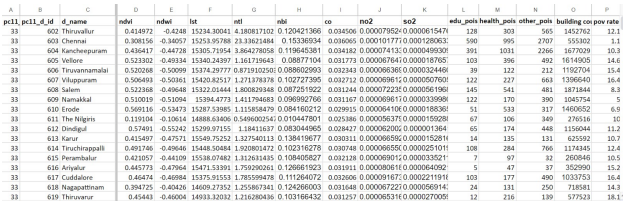
Figure 15. Point Of Interest data for Kanchipuram

Figure 16 shows the buildings filtered based on confidence level from Google Earth Engine Open Building dataset. The green areas represent houses and the count of houses is eventually taken using this for each region.

Figure 16. Building count data

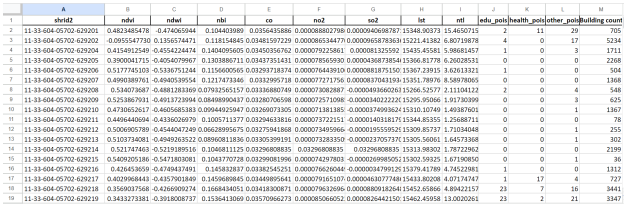
The extracted features are then stored in a file - one for all the districts of Tamil Nadu as shown in Figure 17, another for all the villages/towns for various districts mentioned previously as shown in Figure 18.

The obtained features are used to train the 4 models as mentioned earlier. Each model is trained and tested and the corresponding RMSE, MAE and R2 score are calculated with the help of Equations 4, 5 and 6 respectively. The results are tabulated and shown in Table 1.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
pc11_	pc11_d_id	d_name	ndvi	ndwi	lst	ntl	nbi	co	no2	so2	edu_pois	health_pois	other_pois	building cou	pov rate
33	602	Thiruvallur	0.414972	-0.4248	15234.30041	4.180817102	0.120421366	0.034506	0.000079524	0.000061547	128	303	565	1452762	12.1
33	603	Chennai	0.308156	-0.34057	15253.95788	23.33621484	0.15336934	0.036065	0.000101777	0.000128063	590	995	2707	555302	1.1
33	604	Kancheepuram	0.436417	-0.44728	15305.71954	3.864278058	0.119645381	0.034182	0.000074153	0.000049930	391	1031	2266	1677029	10.3
33	605	Vellore	0.523302	-0.49334	15340.24397	1.161719643	0.08877104	0.031773	0.000067641	0.000018765	103	396	492	1614905	14.6
33	606	Tiruvannamalai	0.520268	-0.50099	15374.29777	0.871910250	0.088602993	0.032343	0.000066366	0.000032446	39	122	212	1192704	15.4
33	607	Viluppuram	0.506493	-0.50361	15420.82517	1.271378378	0.102727395	0.032712	0.000069615	0.000050760	122	227	663	1396640	16.4
33	608	Salem	0.522368	-0.49648	15322.01444	1.800829348	0.087251922	0.031244	0.000072237	0.000056196	145	541	481	1871844	8.3
33	609	Namakkal	0.510019	-0.51094	15394.4773	1.411794683	0.096992766	0.031167	0.000069611	0.000033998	122	170	390	1045754	5
33	610	Erode	0.569116	-0.53473	15287.53985	1.115858479	0.084160212	0.029915	0.000064106	0.000018636	51	533	317	1460652	6.9
33	611	The Nilgiris	0.119104	-0.10614	14888.63406	0.549600254	0.010447801	0.025386	0.000056371	0.000015928	67	106	349	276516	10
33	612	Dindigul	0.57491	-0.55242	15299.97155	1.18411637	0.083044965	0.028427	0.000062002	0.000001364	65	174	448	1156044	11.2
33	613	Karur	0.415497	-0.47571	15549.75252	1.327540113	0.138419677	0.030311	0.000066596	0.000015281	14	135	131	625592	10.7
33	614	Tiruchirappalli	0.491746	-0.49646	15448.50484	1.920801472	0.102316278	0.030748	0.000066556	0.000025101	108	284	766	1174345	12.4
33	615	Perambalur	0.421057	-0.44109	15538.07482	1.312631435	0.108405827	0.032128	0.000069012	0.000033521	7	97	32	260846	10.5
33	616	Ariyalur	0.445773	-0.47964	15471.53991	1.759290261	0.126661923	0.031911	0.000080611	0.000064092	5	47	37	352990	15.2
33	617	Cuddalore	0.46474	-0.46984	15375.91553	1.785599478	0.111264072	0.032606	0.000091671	0.000221191	103	177	490	1033753	16.4
33	618	Nagapattinam	0.394725	-0.40426	14609.27352	1.255867341	0.124266003	0.031648	0.000067227	0.000056914	24	131	250	718581	14.3
33	619	Thiruvarur	0.45443	-0.46004	14933.32032	1.216280436	0.103166432	0.031257	0.000065316	0.000027005	12	216	139	577523	18.1

Figure 17. Zonal level extracted features for Tamil Nadu



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	shrid2	ndvi	ndwi	nbi	co	no2	so2	lst	ntl	edu_pois	health_pois	other_pois	Building count
2	11-33-604-05702-629201	0.4823485478	-0.474065944	0.104403989	0.035643586	0.000088802798	0.000094698787	15348.90373	15.4650715	2	11	29	705
3	11-33-604-05702-629202	-0.095554730	0.135657447	0.118154845	0.034815972	0.000086334477	0.000096587836	15221.41382	6.807198	4	0	17	5234
4	11-33-604-05702-629204	0.4154912549	-0.455422447	0.140409560	0.034503567	0.000079225861	0.000081325592	15435.45581	5.986814	1	0	3	1711
5	11-33-604-05702-629205	0.3900041715	-0.405407996	0.130388671	0.034373514	0.000078565930	0.000043687385	15366.81778	6.26028531	0	0	0	2268
6	11-33-604-05702-629206	0.5177745109	-0.533675124	0.115660556	0.032937183	0.000076443910	0.000083875150	15367.23915	3.626133211	1	0	0	504
7	11-33-604-05702-629207	0.4990389761	-0.490539554	0.121747334	0.033299571	0.000077271756	0.000083704319	15351.78976	8.58978065	0	0	0	1368
8	11-33-604-05702-629208	0.534073687	-0.488128369	0.079325651	0.033368074	0.000073082887	-0.000049366026	15266.52577	2.11104122	2	0	4	548
9	11-33-604-05702-629209	0.5253867931	-0.491372994	0.084989904	0.032807065	0.000072571098	-0.000034022222	15295.95066	1.91730399	0	0	3	625
10	11-33-604-05702-629210	0.4730652617	-0.460568383	0.099449259	0.032690073	0.000071381385	-0.000037499362	15310.10749	1.49387601	0	0	1	1367
11	11-33-604-05702-629211	0.4496440694	-0.433602697	0.100571377	0.032946338	0.000073722151	-0.000014031817	15344.85355	1.25688711	0	0	0	78
12	11-33-604-05702-629212	0.5006905789	-0.454404724	0.066289956	0.032759418	0.000073495966	-0.000019555952	15309.85737	1.71034048	0	0	1	255
13	11-33-604-05702-629213	0.5103734081	-0.494926532	0.089608118	0.032953991	0.000073283350	-0.000023705737	15305.56061	1.64573368	0	0	1	302
14	11-33-604-05702-629214	0.52174763	-0.521189516	0.104681125	0.032960883	0.03296808835	0.03296808835	15313.98302	1.78722962	0	0	0	2199
15	11-33-604-05702-629215	0.5409205186	-0.547180381	0.104377072	0.032990819	0.000074297803	-0.000026985052	15302.59325	1.67190850	0	0	1	36
16	11-33-604-05702-629216	0.426453059	-0.479434749	0.145832837	0.033825425	0.000076620448	-0.000034799126	15379.41789	4.74522881	1	0	0	1312
17	11-33-604-05702-629217	0.4029968443	-0.435701849	0.145968985	0.034498956	0.000079165107	0.000046307774	15433.80208	4.07174747	1	17	4	727
18	11-33-604-05702-629218	0.3569037568	-0.426690274	0.166843405	0.034183008	0.000079632696	0.000080918264	15452.65866	4.89422157	23	7	16	3441
19	11-33-604-05702-629219	0.3433273381	-0.391808737	0.153641306	0.035709662	0.000082664215	0.000082664215	15462.45958	13.00202661	23	2	21	3347

Figure 18. Zonal level extracted features for Kanchipuram

$$RMSE(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}} \tag{4}$$

$$MAE(y, \hat{y}) = \frac{\sum_{i=0}^{N-1}|y_i - \hat{y}_i|}{N} \tag{5}$$

$$R2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{6}$$

Zonal level	Model	RMSE	MAE	R2 score
Village/Town	DTR	11.843	7.944	0.699
	RFR	10.178	7.077	0.778
	CNN	10.701	7.552	0.755
	MLP	10.832	7.604	0.749
District	RFR	1.446	1.202	0.514

Table 1. Evaluation metrics

On comparison, the Random Forest Regressor at the village level resulted in a better R2 score which is 0.778 and the error rates RMSE and MAE are 10.178 and 7.077 respectively which are less compared to others. Hence, this model is chosen as the best one and proceeded with the next step - poverty map generation at the village/town level. At the district level, Random Forest Regressor produced an R2 score of 0.514, MAE of 1.202 and RMSE of 1.446.

Random Forest Regressor outperforms other models due to its ability to handle complex relationships, robustness to overfitting, ability to handle non-linear relationships,it’s scalability and the ability to reduce the variance. By aggregating the predictions of multiple trees, Random Forests can capture complex relationships between the input features and the poverty rate which is crucial in the context of poverty prediction, where various socio-economic and environmental factors may influence the poverty level in different regions. Random Forests are

less prone to overfitting compared to individual decision trees which they achieve by using random subsets of features for each tree and averaging their predictions. This helps the model to filter out the noise and focus on the most important predictors of poverty.

Decision trees tend to be less accurate than Random Forests because they can easily overfit the training data, especially in this complex dataset with many features it may struggle to capture the complexity of the relationship between predictors and poverty. While deep learning models like CNNs and MLPs are powerful in handling complex patterns in data, they may require a vast amount of data to train effectively. Additionally, they might not perform as well as Random Forests when the relationships between input features and the target variable are not inherently spatial or hierarchical, as in the case of poverty prediction where various socio-economic and environmental factors play a role.

The poverty maps for Tamil Nadu and the districts - Kanchipuram, Nawada, Coimbatore and Purulia generated with the help of QGIS software using the predicted values obtained with the help of Random Forest Regressor.

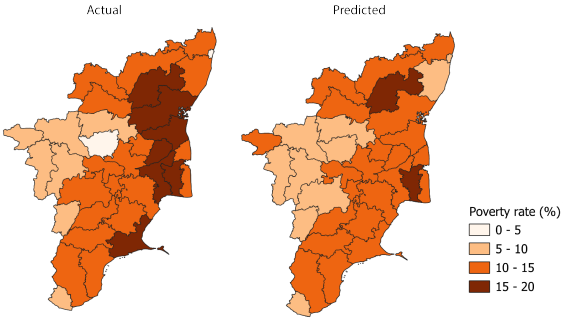


Figure 19. District level poverty map of Tamil Nadu

Figure 19 shows the actual and predicted poverty rates for each district with the help of a random forest regressor model for Tamil Nadu. As shown in the Figure 19, the poverty rates in Tamil Nadu’s districts are less than 20%.

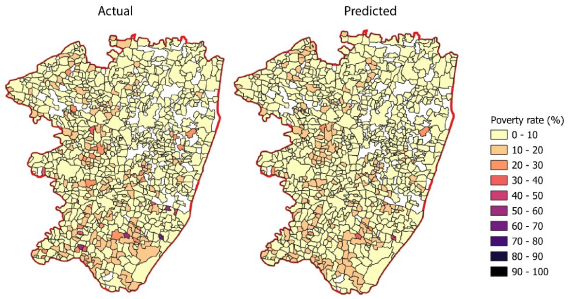


Figure 20. Village/town level poverty map of Kanchipuram

Figure 20 shows the actual and predicted poverty rates with the help of a random forest regressor model for Kanchipuram The overall poverty in Kanchipuram is comparatively less. Kanchipuram is a well-developed district in Tamil Nadu.

Nawada is a district in Bihar, India. Nawada, like many other districts in Bihar and across India, has faced many challenges

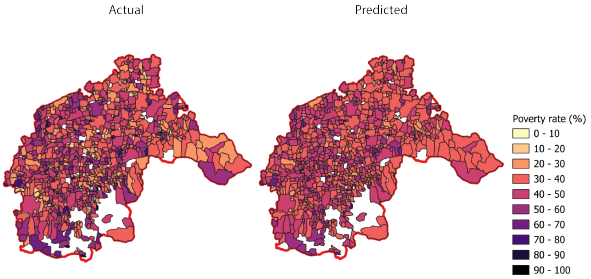


Figure 21. Village/town level poverty map of Nawada

related to poverty and underdevelopment. According to SECC, Nawada has a poverty rate of 52.8% while Kanchipuram has a poverty rate of 10.3% in the year 2011. Limited access to quality education, healthcare facilities, and basic infrastructure like roads and electricity contributes to the perpetuation of poverty. The lack of industries and job opportunities within the district often leads to unemployment or underemployment among the local population. Figure 21 depicts the poverty map of Nawada at the village/town level. The predicted poverty rates of villages and towns in Nawada are high when compared to Kanchipuram.

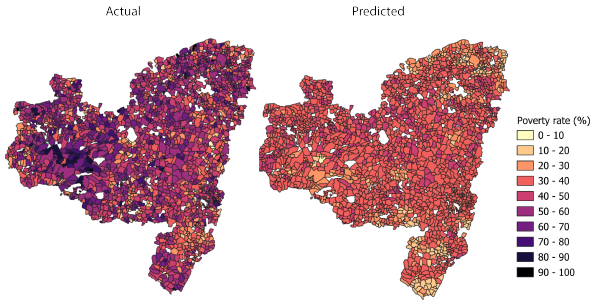


Figure 22. Village/town level poverty map of Purulia

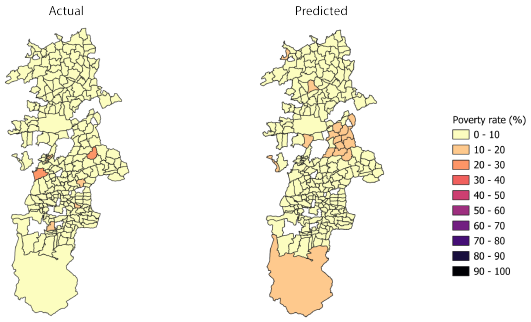


Figure 23. Village/town level poverty map of Coimbatore

Two districts - Purulia in West Bengal and Coimbatore in Tamil Nadu are tested at the village/town level to check how generalized the model is. Figure 22 shows the visualization of the poverty map of Purulia and Figure 23 visualizes the poverty map of Coimbatore region at village/town level.

5. Conclusion

The study makes use of multi-source data from satllites and geospatial systems to predict poverty rate at district and vil-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-2024-319-2024 | © Author(s) 2024. CC BY 4.0 License.

lage/town levels across India. Different environmental and socio-economic factors have been used. Machine learning and deep learning techniques have been deployed to get the best results.

The system was developed using tools and technologies such as GEE API, OpenStreetMap API, Google Colaboratory Notebook, and QGIS. The satellite images are collected with the help of GEE API. The environmental features like NDVI, NDWI, NBI, CO, NO2, SO2, LST and NTL are extracted from the satellite image data. Point of Interest data is collected with the help of osmnx package which is used to retrieve data from OpenStreetMap. The count of education related, health related and other POIs is collected and filtered. Count of building is also one of the features that is used.

QGIS is used to work with shape files for getting geometries of region of interest. The features are extracted for each zone. This data is split for training and testing. 4 different models were used. It is found that the Random Forest Regressor model has an R2 score of 0.778. Using this, we can estimate the poverty in various regions and identify the most poor villages. The policy makers can use this to frame necessary policies for the development of such areas.

References

Aprianto, K., Wijayanto, A. W., Pramana, S., 2022. Deep learning approach using satellite imagery data for poverty analysis in banten, indonesia. *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, 126–131.

Asher, S., Lunt, T., Matsuura, R., Novosad, P., 2021. Development research at high geographic resolution: an analysis of night-lights, firms, and poverty in India using the shrug open data platform. *The World Bank Economic Review*, 35(4).

Avtar, R., Aggarwal, R., Kharrazi, A., Kumar, P., Kurniawan, T., 2019. Utilizing geospatial information to implement SDGs and monitor their Progress. *Environmental Monitoring and Assessment*, 192(1), 35.

Ayush, K., Uzkent, B., Burke, M., Lobell, D., Ermon, S., 2020. Generating interpretable poverty maps using object detection in satellite images. *Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, abs/2002.01612, 4367–4373.

Babenko, B., Hersh, J., Newhouse, D., Ramakrishnan, A., Swartz, T., 2017. Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in mexico. *31st Conference on Neural Information Processing Systems*.

Hu, S., Ge, Y., Liu, M., Ren, Z., Zhang, X., 2022. Village-level poverty identification using machine learning, high-resolution images, and geospatial data. *International Journal of Applied Earth Observation and Geoinformation*, 107, 102694. doi.org/10.1016/j.jag.2022.102694.

Jean, N., Burke, M., Xie, M., Davis, W., Lobell, D., Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), 790-794.

Lin Htet, N., Kongprawechnon, W., Thajchayapong, S., Isshiki, T., 2021. Machine learning approach with multiple open-source data for mapping and prediction of poverty in myanmar. *2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 1041–1045.

Luo, E., Kuffer, M., Wang, J., 2022. Urban poverty maps - From characterising deprivation using geo-spatial data to capturing deprivation from space. *Sustainable Cities and Society*, 84, 104033. doi.org/10.1016/j.scs.2022.104033.

Ministry of Rural Development, 2011. Socio Economic and Caste Census. Ministry of Rural Development, Government of India. secc.gov.in.

Putri, S. R., Wijayanto, A. W., Pramana, S., 2023. Multi-source satellite imagery and point of interest data for poverty mapping in East Java, Indonesia: Machine learning and deep learning approaches. *Remote Sensing Applications: Society and Environment*, 29, 100889. doi.org/10.1016/j.rsase.2022.100889.

Puttanapong, N., Martinez, A., Bulan, J., Addawe, M., Durante, R., Martillan, M., 2022. Predicting Poverty Using Geospatial Data in Thailand. *ISPRS International Journal of Geo-Information*, 11(5), 293.

Tingzon, I., Orden, A., Go, K. T., Sy, S., Sekara, V., Weber, I., Fatehkia, M., García-Herranz, M., Kim, D., 2019. Mapping poverty in the Philippines using machine learning, satellite imagery, and crowd-sourced geospatial information. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-4/W19, 425–431. doi.org/10.5194/isprs-archives-XLII-4-W19-425-2019.