OPTIMIZING SOLAR PANEL PLACEMENT

A MACHINE LEARNING APPROACH TO MAPPING SOLAR ENERGY POTENTIAL

MODULE BUSM: 130 GROUP PROJECT

Supervisor: Natalia Efremova, Meghna Asthana

Programme: Business Analytics MSc – 2023/24

Ankit Mate (230256765)
Himanshu Sharma (230369788)
Parth Singh (221150274)
Ramsha Najam Kazi (230230444)
Vatsal Vipul Doshi (230504415)

Queen Mary University of London

Student's Declaration

We,

Ankit Mate, Student ID: 230256765

Himanshu Sharma, Student ID: 230369788

Parth Singh, Student ID: 221150274

Ramsha Najam Kazi: Student ID: 230230444

Vatsal Vipul Doshi, Student ID: 230504415

hereby declare that the work in this Group Project Report is our original work. We have not copied from any other students' work, work of ours submitted elsewhere, or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written for us by another person. The individual sections of the report have been written by the author in the section title with no external assistance.

Queen Mary University of London

Contents

Queen Mary University of London

Executive Summary (Group 3)

The attached report explores the development of a comprehensive model for mapping solar energy potential, with a specific focus on the Cambridge area in the United Kingdom. This project aims to address the critical business challenge of identifying optimal locations for solar panel installations, considering various geographical and environmental factors.

As the world shifts towards renewable energy sources to combat climate change, solar energy has emerged as a key player. However, the efficiency and economic viability of solar installations heavily depend on their geographical location. Factors such as direct sunlight exposure, seasonal variations, and local climate conditions significantly impact the potential for solar energy generation. Incorrect positioning of solar panels can lead to substantial financial losses and inefficient use of resources. Therefore, accurate mapping of solar energy potential is crucial for investors, energy companies, and policymakers to make informed decisions about solar infrastructure development.

The project utilizes a range of data sources and analytical techniques:

1. **Weather Data:** UK MET office weather data, including temperature, rainfall, and sunshine duration.
2. **Satellite Imagery:** Sentinel-2 satellite data, providing high-resolution spectral information.
3. **Digital Elevation Model (DEM):** SRTM (Shuttle Radar Topography Mission) data for terrain analysis.

The methodology involves several key steps:

1. **Data Processing:** Cleaning and integrating weather data, satellite imagery, and elevation data.
2. **Spectral Analysis:** Calculation of indices like NDVI (Normalized Difference Vegetation Index) and NDBI (Normalized Difference Built-up Index) from satellite bands.
3. **Terrain Analysis:** Slope calculation from DEM data.
4. **Machine Learning Models:** Implementation of various models including Random Forest, Gradient Boosting, XGBoost, Artificial Neural Network (ANN), and Deep Neural Network (DNN) for predicting solar potential.

The project employed several sophisticated analytical techniques to process and analyse the diverse datasets:

1. Sentinel-2 Spectral Bands Analysis:

The study utilized multiple spectral bands from Sentinel-2 satellite imagery, including:

- **Blue Band (Band 2):** Used for coastal and water body monitoring, and analysing vegetation and urban areas.
- **Green Band (Band 3):** Critical for vegetation analysis and detecting plant health.
- **Red Band (Band 4):** Essential for vegetation analysis and calculating NDVI.
- **Near-Infrared Band (Band 8):** Crucial for vegetation health analysis and land cover classification.
- **Short-Wave Infrared Band (Band 11):** Useful for detecting moisture content in soil and vegetation.

These bands were processed using the rasterio library in Python, allowing for efficient reading and manipulation of geospatial raster data.

2. NDVI (Normalized Difference Vegetation Index) Calculation:

NDVI was calculated using the formula: NDVI = (NIR - RED) / (NIR + RED)

This index is crucial for identifying areas with healthy vegetation, which are generally less suitable for solar panel installation.

3. NDBI (Normalized Difference Built-up Index) Calculation:

This index helps in identifying built-up or urbanized areas, which can be potential locations for rooftop solar installations.

4. Slope Calculation from DEM:

The slope was derived from the Digital Elevation Model using gradient calculations in both x and y directions.

5. Otsu's Thresholding Method:

This unsupervised method was used to automatically determine optimal threshold values for separating different land cover types in the absence of ground truth data.

Machine Learning Model Development:

The project explored various machine learning models to predict solar energy potential:

1. Random Forest Regressor. It advantages are that it handled non-linear relationships well, resistant to overfitting. Implementation: Used raw feature values instead of predefined thresholds, allowing the model to learn relationships from data.

2. XGBoost Regressor: It is Known for efficiency and high performance in structured data. The model is compared using Mean Squared Error (MSE) and R-squared (R^2) scores

3. Gradient Boosting Regressor: It iteratively improves predictions by combining weak models. The True vs. Predicted plots were created for performance analysis.

4. Artificial Neural Network (ANN): In ANN, the features were scaled using StandardScaler. Three hidden layers were used. ANN achieved the best results with lowest MSE (0.000045) and highest R^2 (0.997).

5. Deep Neural Network (DNN): Five hidden layers for capturing more complex patterns. It was the second-best model with MSE of 0.000059 and R^2 of 0.996.

Model Training and Evaluation Process:

1. Data Splitting: The dataset was split into 70% training and 30% testing sets to ensure unbiased evaluation.

2. Feature Preparation: Included spectral bands, NDVI, NDBI, elevation data, and slope.

3. Target Variable: A complex target variable was created by combining factors like NDVI, slope, elevation, and land cover classification.

4. Model Training: Each model was trained on the prepared features to predict the solar potential score.

5. Evaluation Metrics: Mean Squared Error (MSE) and R-squared (R^2) were used to assess model performance.

6. Visualization: Predicted solar potential was visualized on maps for each model, allowing for easy comparison and interpretation.

Insights from Model Performance:

Neural Network Superiority: Both ANN and DNN significantly outperformed other models, suggesting that the complex, non-linear relationships in the data are best captured by these architectures.

Gradient Boosting Limitations: While still performing well, the Gradient Boosting model showed higher prediction errors, indicating it might not capture some of the nuanced patterns in the data as effectively as the neural networks.

Random Forest Performance: The Random Forest model performed admirably, suggesting that ensemble methods are well-suited for this type of geospatial prediction task.

Model Consistency: The consistency in predictions across different models, especially in identifying high-potential areas, reinforces the reliability of the results.

Practical Applications and Business Impact:

1. Investment Decision Support: The high-accuracy predictions from the ANN and DNN models can significantly reduce investment risks in solar energy projects. Investors can use these maps to identify prime locations for solar farms or large-scale installations, potentially increasing ROI and reducing payback periods.

2. Urban Planning and Development: Urban planners can utilize the solar potential maps to: Integrate solar energy considerations into new development projects, identify optimal locations for community solar projects, and Plan 'solar-ready' zones in cities, encouraging sustainable urban development.

3. Policy Formulation: Policymakers can use these insights to: Develop targeted incentives for high-potential areas, create zoning laws that encourage solar adoption in optimal locations, set realistic and achievable renewable energy targets based on regional potential.

Grid Infrastructure Planning: Energy companies and grid operators can use this information to plan grid reinforcements in areas with high solar potential, develop strategies for integrating distributed solar generation into the existing grid, optimize the placement of energy storage systems to complement solar installations.

4. Environmental Impact Assessment: The outputs can assist in identifying areas where solar installations would have minimal environmental impact, balancing land use between solar energy generation and other purposes (e.g., agriculture, conservation).

Challenges and Future Directions:

1. Dynamic Environmental Factors: Future models could incorporate more dynamic factors like seasonal variations in weather patterns and long-term climate change projections.

2. Integration with Other Renewable Sources: Expanding the model to include potential for other renewable sources (wind, hydroelectric) could provide a more comprehensive energy planning tool.

3. Economic Factors Integration: Incorporating economic data such as land prices, construction costs, and local electricity prices could enhance the model's practical applicability.

4. Real-time Data Integration: Developing systems to continuously update the model with real-time satellite and weather data could provide more dynamic and current predictions.

5. Scalability and Generalization: Further research is needed to ensure the model's applicability across diverse geographical regions with different climatic and topographic characteristics.

Conclusion:

This project demonstrates the power of combining advanced data analytics, machine learning, and geospatial technology to solve complex problems in renewable energy planning. The developed models, particularly the ANN and DNN, show remarkable accuracy in predicting solar energy potential, providing a valuable tool for decision-makers across various sectors.

The methodology and findings extend beyond solar energy mapping and have potential applications in various fields requiring geospatial analysis and predictive modelling. As the world continues to grapple with the challenges of climate change and the transition to renewable energy, tools, and methodologies like those developed in this project will play a crucial role in shaping sustainable and efficient energy landscapes.

The success of this project underscores the importance of interdisciplinary approaches in tackling complex real-world problems. By leveraging diverse data sources, advanced analytical techniques, and machine learning, we can gain insights that were previously unattainable, paving the way for more informed decision-making and strategic planning in the renewable energy sector and beyond.

The ANN and DNN models demonstrated excellent predictive accuracy with minimal errors, explaining over 99% of the variance in the data. These models provide highly reliable predictions for solar potential across the Cambridge region.

Visualization and Interpretation:

The report includes visualizations of solar potential maps generated by different models. These maps highlight areas with varying degrees of solar energy potential, with brighter (yellow) areas representing higher potential and darker (purple) areas indicating lower potential. The consistency across different model outputs reinforces the reliability of the predictions.

Challenges and Limitations:

The project faced several challenges:

- Data Integration: Merging diverse data sources with different formats and resolutions.
- Computational Demands: Processing large volumes of satellite and geographical data.
- Model Accuracy: Ensuring reliability and avoiding overfitting in predictive models.
- Scalability: Developing models that can be applied to different geographical regions.

Recommendations for Stakeholders:

Based on the analysis, the report offers tailored recommendations for different stakeholder groups:

Policymakers:

Maintain and enhance incentives for solar energy investment in high-potential areas.

Update zoning regulations to encourage solar panel installations.

Align long-term renewable energy goals with identified high-potential zones.

Invest in storage solutions and grid upgrades to support increased solar energy integration.

Urban Planners:

Develop community solar projects in underutilized urban spaces.

Integrate solar energy into backup power strategies for critical infrastructure

Investors:

Prioritize investments in regions with the highest predicted solar potential.

Diversify solar portfolios across different high-potential zones.

Consider long-term investments in large-scale solar farms in rural high-potential areas.

Explore partnerships with local governments and businesses for co-development of solar projects.

Future Explorations and Improvements:

The report suggests several avenues for further research and model enhancement:

Model Optimization: Fine-tuning hyperparameters and exploring ensemble methods to improve predictive performance.

Additional Data Integration: Incorporating more diverse datasets such as historical trends, demographic data, and detailed weather information.

Feature Engineering: Creating new features or transforming existing ones to capture complex patterns in the data.

Transfer Learning: Applying pre-trained models from similar tasks to reduce training time and improve performance.

This project demonstrates the power of integrating diverse data sources and advanced machine learning techniques to solve complex business problems in the renewable energy sector. The developed models, particularly the ANN and DNN, show excellent potential for accurately predicting optimal locations for solar panel installations. These insights can significantly impact decision-making processes for policymakers, urban planners, and investors in the solar energy sector.

The methodology and findings of this project extend beyond solar energy mapping and can be applied to a wide range of business problems across various industries. From optimizing energy usage and managing financial risks to improving retail operations and healthcare outcomes, the techniques explored in this project have broad commercial applications.

As the world continues to shift towards renewable energy sources, the importance of accurate solar potential mapping will only grow. This project provides a robust framework for making data-driven decisions in solar energy infrastructure development, contributing to more efficient resource allocation, reduced financial risks, and accelerated progress towards sustainable energy goals.

Future research should focus on ways to accelerate domestic investment in renewable energy and storage solutions. It should also assess the impact of a rapid energy transition on service and resource demands, including the need for skilled labor and ethical sourcing of components. With resources, technology, and costs no longer being significant barriers, the focus should now shift to addressing the political and economic challenges of an accelerated energy transition.

By leveraging the insights and methodologies presented in this report, stakeholders can make more informed decisions, optimize their solar energy strategies, and contribute to the broader goal of creating a sustainable and efficient renewable energy infrastructure.

Business Problem (Parth Singh)

Over the world there has been a change in energy demand and supply due to the strategies that are being put in place to use renewable energy to reduce on emissions of greenhouse gases. An important aspect of the above changes is solar energy as a reliable, available and less costly form of energy as compared to traditional energy. The uptake of solar energy is regarded as critical in the quest for unlimited access to clean energy as nations seek to ignite the lit tires of the globally recognized Paris Agreement (IEA, 2021). However, no less significant is the question of where solar energy should be generated – and this question remains one of the greatest challenges in the future of solar energy. The primary organizational problem that this project seeks to address is this. This means that factors such as geographical location, meteorological conditions as well as seasonal variations are just but a few of the many factors that define the potential of solar energy generation. Thus, while some regions have got equal access to as much direct sunlight for a year as any regions in equatorial regions have, many other regions could face severe fluctuations in direct sunlight due to geographical conditions, cloud coverage and sky conditions for the year (U. S. Department of Energy, 2020). But given this rationale, it is unfitting to install solar panels in a uniform way. However, in order to ensure that investments are made in solar structures that are both economically and effective, it is crucial that the mapping of the solar energy potential of the various places is accurately accomplished. The issue becomes more complicated when the stakes of the economy and environment join the stakes of business. Incorrect positioning or poor orientation of the solar panels is one of the greatest mistakes you can make since it can cost you a fortune. Sectors with low solar potential may result to low energy production, long time cycle to recover investment and hence low return on investment if investment is made on that site (World Bank, 2019). They may also distort the goals of sustainable development from the environmental perspective because it is associated with wasteful consumption of resources and land. Thus, precise mapping of solar energy potential is not only a matter of technology but also a vital business requirement. In addition, the recent advancement in data-driven technologies has forced the use of satellite imaging and meteorological information which are instruments in developing comprehensive models for solar potential assessment (Ghafoor & Munir, 2021).

These models can hence provide the required data to the stakeholders, including governments, investors, and energy companies, on where to install the solar panels so that maximum energy can be generated and financial losses minimized. However, building such models is not an easy task, that is using, storing, sorting and analysing of the great amount of data, and also ensuring the accuracy and reliability of the predictions. As such, the need to come up with an elaborate model that will capture the viability of solar energy based on geography is the business problem this project aims to solve. This model aims at predicting the best areas for putting up the solar panels using satellite data and climatological data. This will enhance the rate of return of solar investments and the more general goal of sustainable development of energy systems. The Need for Accurate Solar Energy Potential Mapping There's increasing need for planning to incorporate solar energy system as the world shifts towards renewable energy system. They specifically stress the role of a correct identification of places which can potentially produce the largest amounts of solar energy. This condition stems from variations in inherent characteristics of radiation of the sun and effects of geographical and climatic features on solar power generation (U. S. Department of Energy, 2020). Hence, the accurate geographical description of the solar energy resource is not only the technical problem but may be considered as a critical strategic business need that has far reached implications for energy security, environmental responsibility, and profitability. At this point, it will be important to explain why it is possible to speak about the definitive importance of adequate solar energy potential mapping. Their usefulness is major in the aspect of supporting decision-making process hence desirable for efficient management of resources in solar energy. Stakeholders can determine the feasibility of solar power in a given area with a lot of accuracy through the use of models that incorporate images from space and climatic data (Ghafoor & Munir, 2021). This strategy is in contrary to the practices that often involve estimation that are gross and less accurate when compared to the present data points; such broad estimations fail to capture regional variations in solar radiation. This ensures that the solar panels are effectively and optimally situated to generate maximum energy or electricity which in overall boosts the overall profitability and effectiveness of solar energy plants. Accurate identification and mapping of solar energy potential has considerable consequences on the economic perspective of a country. Ongoing capital investments are required for solar energy projects and, depending on how much power they generate, such projects will generate profits. In the solar projects classification depending on power output, the World Bank (2019) mentions that poor site selection choice causes low performance of the installations and impacts the return on investment and productivity of the solar paybacks. For instance, energy generated by solar farm may fail to generate enough revenue to cater for its installation as well as maintenance

when located in a region with little exposure to sunlight. But mistakes are usually avoided by ensuring that allocations match solar plains hence providing steadier and more durable returns. It is also important to map the solar energy potential in the right way from an environmental perspective for integrating sustainability. Another way of reducing the total environmental impact is by placing the solar panel systems strategically where they are likely to receive the greatest levels of direct illumination to yield the most electricity while occupying a very small amount of land (IEA, 2021). In the same respect, accurate mapping contributes to the broader goal of reducing greenhouse emissions and arresting climate change, by harnessing the potential of the solar systems. When possible locations are selected with optimum solar exposure, then the generation of energy through solar power is sustainable, friendly to environment hence the benefits of solar energy are fully achieved without any compromise. Furthermore, there is a strong enhancement in the topic by integrating information on weather and satellite imagery in the generation of maps of solar energy. Solar mapping also requires up-to-date data in respect of the key variables, such as solar radiation, cloud coverage and surface temperature, which are all obtainable from satellite information (Ghafoor, & Munir, 2021). This data can be used in generating accurate models of how much electricity can be generated from solar energy, as long as it is used in conjunction with the ground-based weather data. Due to this, these models are widely useful to energy planners and investors because they provide a bearing for determining where to situate solar power systems. Thus the justification to accurate mapping of solar energy potential on the basis of its critical role played in unearthing the resource efficiency, adding more financial value and promoting ecological responsibility. This way stakeholders may enhance efficiency and outcomes of the solar energy projects and, in the long run, advance the prospects of the global sustainable energy future by employing advanced technological tools to map and predict the solar potential around the world. Geographic and Climatic Variability Impact Seasonal variations, latitude and longitude related factors and other climatological factors such as climate and weather conditions greatly influence the variability in irradiation across various locations and regions, therefore, has a strong correlation with the viability of solar energy conversion. As all these variances determine the efficiency and the total output and flows of solar panels they become crucial for understanding in order the solar energy infrastructure to be efficiently established and installed. The requirement for accurate and thorough mapping to guarantee that solar systems are placed in locations where they can function at their best is highlighted by the geographic and climatic variations in solar energy potential. One of the major factors that determine the amount of Solar Radiation that hits a particular region is its geographical position. Region in closer proximately to the equator therefore receives a direct and higher solar radiation throughout the year (U. S.

Department of Energy, 2020). There, temperature also exhibits large seasonal ranges as do the solar insolation which is longer in summer and shorter in winter compared to the higher latitudes. Due to such uncertainty, it is important to conduct proper assessment of the potential of a certain area in terms of harnessing solar energy before any investments have been made on the needed structures. Even more so, topographical characteristics complicate the assessment of the prospective of solar energy. They pointed out that the quantity of sunlight that can get to the surface could be reduced by shade which may be occasioned by mountains, valleys, as well as other features of the land (NREL, 2018). Besides, the position of the land, that is, whether it faces north, south, east, or west or whether it has a relatively flat surface or a sloping one also determines the amount of heat energy from the sun that it gains. For instance, in locations such as the Northern Hemisphere, the southern-facing slopes see more sun, therefore, are ideal for solar power generation. Thus, topography should be taken into consideration during the process of mapping of the solar potential of the territory as it influences the efficiency of the action of solar panels to a vast extent. Climate is also known to affect the generation of solar power in a considerable way. For example, cloud affect the amount of sunlight that can be collected by the solar panels by greatly reducing the amount that reaches the surface of the earth (World Bank, 2019). Thus, regions with much continuous cloud cover might seem to possess lesser solar possibilities irrespective of the latitudes in which they are located. Like humidity also affects the amount of sun light which is available to be captured by solar panels and air pollution also affects the amount of sun light which can be captured by changing its direction and scattering the energy (Ghafoor & Munir, 2021). For a more accurate assessment of a region's suitability for solar energy projects, these considerations make it essential to include climate data in solar energy potential mapping. Seasonal variations are also very crucial for the development of solar energy production. Unpredictable variations in the amount of energy produced over a given period can occur because of the variation of solar radiation in some places between seasons, such as the summer and winter (U. S. Department of Energy, 2020). Because of this, there is a volatile pattern in designing the solar energy projects because it involves a seasonal pattern especially in regions where there is constant demand for electricity. Perhaps, the design of such solar systems that would be closer to the energy needs of the region can be informed by better mapping that considers these variations across the seasons. The Role of Satellite Imagery and Data Integration Satellite imagery and data integration are critical to ascertain the exact area’s solar potential in the process of enhancing the production of the solar energy continually. These new generation technologies provide the comprehensive and real time view of the surface of the earth thereby providing the greater insight into the factors that influence the generation of solar energy. There is absolutely

no doubt that stakeholders can develop highly accurate models that dictate the right placement of solar panels hence enhancing the prospects of solar energy, business and profitability by merely harnessing satellite imagery in conjunction with ground meteorological data. Remote sensing is the bird's eye view of the Earth that has done paradigm shift in the evaluation of solar power prospect by collecting information regarding a number of factors that influence solar intensity. Photovoltaic measurements and other parameters necessary for solar energy generation such as solar radiation, cloud cover, surface albedo and others are fitted in today's satellites. This data provides a continuous and constant coverage of surface characteristics of the Earth which makes it ideal for evaluating solar resources. Compared to other approaches for ground-based measurements, which can often be time and spatial limited, satellite photography is a powerful tool for assessing solar energy since coverage and change detection can be performed instantly across the region of interest. Satellite large-scale solar potential models are enhanced significantly when complemented with ground meteorological data. When combined with ground data, global or satellite data contain the information about the local state of the atmosphere from the ground-based data such as weather-station measurements of temperature, humidity and wind speed. This provides a more detailed pattern on the regional climate and how it influences the amount of radiation received from the sun (NREL, 2018). For instance, ground sensors give crucial data on the density and more so the movement of the cloud which improves the short-term forecast of the sun's energy even though satellite data may give signals that there are clouds. Due to this integration, it is possible to create reliable and dynamic models that are revived by new data every time it is possible and, thus, provide growing accuracy of potential solar energy estimates (U. S. Department of Energy, 2020). This is especially important in areas with unpredictable climate, so that assessors can modify the energy plans to reflect the current climate for instance, if the solar power generation may reduce because of cloudy weather. Moreover, the satellite data various parameters like vegetation cover and land slope to decide the right place of s Solar Panels. These features are useful for selection of sites that would be most practical and cost-effective (World Bank, 2019). Minimizing the likelihood of choosing suboptimal sites improves the ROI in cases where this information is complemented by the observations of irradiance. However, there are some challenges of which are the significant quantities of processing power required to process large volumes of data obtained from satellites and the need to ensure consistency between satellite and ground pointing measurements to avoid errors (Ghafoor & Munir, 2021). However, these challenges should not be an obstacle that nullifies the benefits that can only make this method crucial for modern Solar energy assessments. Challenges in Developing the Model There is a sheer volume and a formidable level of disaggregation of data that is needed

to underpin a solar energy potential model. A rich amount of data is obtained through geographic information systems (GIS data), satellite imagery, ground-based meteorological data; generated data require post-processing, assessment, and integration into a model (Ghafoor & Munir, 2021). Among the variables in this data, which may widely vary from location to location as well as from time to time, are the land relief, amount of received solar radiation, level of cloud cover, and surface temperature. This requires a lot of computational power, as well as algorithms that can operate and analyze large amounts of data, and come up with meaningful insights. Another challenge is that of ascertaining the correctness and reliability of a model. Owing to the variation of the solar radiation and the weather conditions, even a small data error leads to a big error in the prediction of the model (U. S. Department of Energy, 2020). To avoid such pitfalls, for instance, disparities are likely to be obtained due to differences between the satellite data and ground measurements, data processing has to be done well. This can be managed by developers through the following validation methods: multiple sources data comparisons and checking if the model has the ability to predict past data. However, this procedure is time consuming and requires a good amount of meteorological and data science knowledge. Another worthwhile problem is the issue of merging several sources of data. The properties of each data source, the format, and the constraints that are tied to each data source also vary and range from satellite imaging, ground-based weather observations to geography information system (GIS) data (NREL, 2018). Such masses of data should be appropriate to their compatibility and have no bias when they are integrated into one model. This makes it difficult to develop a model when the factors influencing the potential of solar energy are of dynamic nature. Because of the fluctuating nature of the solar radiation, meteorological and other atmospheric conditions, it is almost impossible to create a model that is correct in all cases at all the time (World Bank, 2019). For these factors, developers have to include variables that provide actual-time data input and refresh the model from time to time. But doing so invokes the state of high-quality hardware and real-time processing efficacy, which can be costly and application intense. And last, one of the main challenges which is usually addressed when designing models is scalability. Because of differences in the geographical and climatic factors, a model that is appropriate in one region may be difficult to implement in another (Ghafoor & Munir, 2021). This means that for the model to be effective, developers ought to see to it that the model is expandable as well as flexible in a way that it will not distort performance whenever it is readjusted to meet regional demands. This entails developing procedures that may be adapted to the prevailing conditions in those areas as well as incorporating data related to those areas. Potential Business Outcomes Excluding revenues which may be generated within a short-term focus, the right development and implementation of the solar mapping potential

model can lead to the following business consequences. One of the main conclusions is the improvement of efficiency of using resources in projects of the usage of solar energy. Investments can then be planned well in areas that will yield the most returns as the different locations are determined to have different amounts of solar capacity (Ghafoor & Munir, 2021). Instead of appearing as a scattering of solar farms all over the world, this pinpoints the level of investment and increases the total profitability and effectiveness of solar energy. Correct geographical identifying of the possible solar energy production not only increases investment, but also significantly reduces costs and therefore the level of potential losses on a solar energy project. Data-backed projects provide higher potential to attract investors as they provide better vision of planned energy production and likely ROC (World Bank, 2019). Hopes are high that a little more investor confidence will reduce the costs of capital even further and improve the financing arrangements for solar projects and thus make them more financially feasible. However, it must also not be forgotten what helpful consequences correct designation of the solar power has for the environment and the community. It assists in optimising energy production, minimal land and space utilization, and reduced environmental impact by placing the solar panels in the most effective locations (U. S. Department of Energy, 2020). This is in accord with the sustainable development goals of the global world and at the same time consolidates the position of solar energy project firms as champions of environmental conscience. In other words, the establishment of an enhanced model of spatial analysis to the solar resource has the potential to enhancement of utility of resource, reduction of cost and financial exposure, and enhanced social and environmental outcomes. All in all, these benefits explain why solar energy protracts is sustainable and plays a significant role to provide renewable energy in the world. Broader Industry Implications and Conclusion The creation and application of an accurate solar energy potential mapping model will have wider effects on the renewable energy sector overall. This model can impact industry best practices and encourage the sector to embrace more data-driven techniques by establishing a new benchmark for accuracy in solar energy site selection (Ghafoor & Munir, 2021). This move toward accuracy and efficiency boosts the profitability of solar installations while also adding to the general appeal and credibility of investments in renewable energy. The sector may witness a marked boost in the scalability and efficiency of solar energy projects as more businesses use these cutting-edge mapping techniques, hastening the world's shift to renewable energy sources (World Bank, 2019). In conclusion, a revolutionary chance to maximize solar energy expenditures is presented by the incorporation of satellite imagery and meteorological data into an all-encompassing solar energy potential mapping model. Through the resolution of issues related to climate and geographic variability, data integration, and model correctness, this

strategy not only optimizes financial gains but also advances more general environmental and social objectives. Globally, more sustainable and effective energy production methods might be encouraged by the wider adoption of such models, which might establish new benchmarks for the renewable energy sector (U.S. Department of Energy, 2020).

Literature Review (Ramsha Najam Kazi)

The shift toward renewable energy has placed solar photovoltaic (PV) systems at the forefront of global energy strategies. Solar PV harnesses the sun's energy to generate electricity, making it an attractive option in the fight against climate change. However, the integration of solar energy into the electricity grid presents unique challenges due to its intermittent nature, primarily influenced by changing weather conditions. Accurate forecasting of solar PV output is crucial for maintaining grid stability and optimizing the management of energy resources.

This review explores the methodologies used to identify the best locations for solar PV installations, focusing on integrating geospatial data such as the Normalized Difference Vegetation Index (NDVI), slope, and elevation. Additionally, the review examines the use of machine learning models to improve the accuracy of solar PV forecasting, as highlighted in the literature.

Geospatial Analysis for Optimal Solar PV Installation

Selecting the most suitable sites for solar PV installations involves considering various environmental and geographical factors. The code developed for this project integrates datasets such as NDVI, slope, elevation, and cloud cover to identify the best locations for solar PV panels. This approach is consistent with recent trends in the literature, which emphasize the importance of using geospatial data in renewable energy planning (Mathe et al., 2024).

NDVI and vegetation considerations

NDVI is an essential tool for filtering out areas with dense vegetation. High NDVI values typically indicate significant vegetation, which can cause shading and reduce the efficiency of solar panels. By applying an NDVI threshold (e.g., NDVI < 0.2), the developed code effectively identifies areas with sparse or no vegetation, making them ideal for solar installations. This approach is supported by studies that underscore the negative impact of vegetation on solar

PV efficiency due to shading and the associated increase in maintenance costs (Mathe et al., 2024).

Slope

The slope of the land is another critical factor in determining the suitability of a site for solar PV installation. Flat terrain is generally preferred because it simplifies construction and usually receives more consistent sunlight. The code calculates slope using Digital Elevation Model (DEM) data, excluding areas with steep slopes (e.g., slope > 15 degrees) that might not be suitable for solar panels. Research by Xia et al. (2024) supports this approach, noting that both slope and aspect significantly influence solar irradiance, with south-facing slopes in the Northern Hemisphere receiving optimal sunlight.

Elevation

While elevation itself is not a primary criterion for solar PV site selection, it can influence local microclimates and, consequently, solar irradiance. Higher elevations may receive more sunlight and less shading from surrounding features. The code takes elevation data into account to indirectly assess these microclimatic factors. Studies such as those by Vandal and Nemani (2023) have demonstrated that incorporating elevation data into solar forecasting models can enhance prediction accuracy, particularly in regions with diverse topography.

Machine Learning in Solar PV Forecasting

The use of machine learning (ML) in solar PV forecasting has become increasingly popular due to its ability to model complex, non-linear relationships between various factors influencing PV output. The integration of ML models into the solar PV site selection process, as illustrated by the Random Forest classifier used in the code, represents a significant advancement in the field.

Random Forest Classifier

The code employs a Random Forest classifier to predict the suitability of locations for solar PV installations based on features like NDVI, slope, and radiance. This model was trained on labeled data, identifying areas as suitable or unsuitable based on predefined thresholds. Ensemble methods such as Random Forest are particularly well-suited for handling large, complex datasets and are known for their robustness (Breiman, 2001). Research supports the

effectiveness of Random Forest classifiers in scenarios where multiple input variables are involved, as they can automatically prioritize the most relevant features for making predictions (Mathe et al., 2024).

Integration of Satellite Data

The use of satellite data in PV forecasting has gained importance as it allows for high-resolution monitoring of weather conditions over large areas. Vandal and Nemani (2023) highlight the application of geostationary satellite data in improving the temporal resolution of weather datasets, which is essential for accurate PV forecasting. Their work demonstrates the potential of deep learning models to interpolate and fuse satellite data, thereby enhancing the accuracy of forecasts, particularly in areas with limited ground-based observations.

Data Gathering and EDA (Vatsal Doshi)

Weather Data

The UK MET office weather data is acquired from Kaggle.com, the MET office (Meteorological office) is the United Kingdom's national weather service which is responsible for providing weather forecasts, warnings and climate monitoring.

We have used this dataset to map and determine which of the stations in the UK have characteristics which will make them better for installing solar panels, these characteristics include areas which receives the maximum temperature, minimum temperature and the areas with maximum sunlight.

The dataset contains the following columns:

- **year**: Year in which the measurements were taken
- **month**: Month in which the measurements were taken
- **tmax**: Mean daily maximum temperature (°C)
- **tmin**: Mean daily minimum temperature (°C)
- **af**: Days of air frost recorded that month (days)
- **rain**: Total rainfall (mm)
- **sun**: Total sunshine duration (hours)
- **station**: Station location where measurement was recorded

initially the dataset did not include the columns for latitude and longitude for the stations which was necessary for determining and mapping the locations in the map.

The latitude and longitude columns were then added to the original dataset by using the 'VLOOKUP' function in Excel by creating a separate sheet listing each station's name along with its corresponding latitude and longitude coordinates. Then by using the VLOOKUP function in the original dataset we could match the station names to their corresponding coordinates.

Summary Statistics

The dataset provided consists of 25,043 rows and includes various climate-related features like temperature, rainfall, sunshine, and geographical coordinates. The dataset includes the following features: year, month, tmax (maximum temperature), tmin (minimum temperature), af (air frost days), rain, sun, station, latitude, and longitude.

breakdown of key statistics for each feature:

- **Year:** Data ranges from 1853 to 2020, with a mean year of approximately 1967.
- **Month:** Data covers all 12 months, with a mean of 6.49 (June).
- **Temperature:**
    - **tmax:** The maximum temperature ranges from -0.7°C to 28.3°C, with an average of 12.69°C.
    - **tmin:** The minimum temperature ranges from -8.6°C to 17.0°C, with an average of 5.87°C.
- **Air Frost (af):** The number of air frost days ranges from 0 to 31 days, with a mean of 3.38 days.
- **Rain:** The rainfall ranges from 0 mm to 568.8 mm, with an average of 74.67 mm.
- **Sunshine (sun):** Sunshine duration varies between 4 and 350 hours, with an average of 116.4 hours.
- **Geographical Coordinates:**
    - **Latitude:** Ranges from 50.22° to 60.13°.
    - **Longitude:** Ranges from -6.86° to 1.75°.

Graphical visualization

Graph representing average minimum temperature

The map provides a visual representation of average minimum temperatures across various weather stations in the United Kingdom. Each station is marked on the map, and a colour gradient from blue (indicating colder temperatures) to red (indicating warmer temperatures) highlights the temperature variations in degrees Celsius (°C).

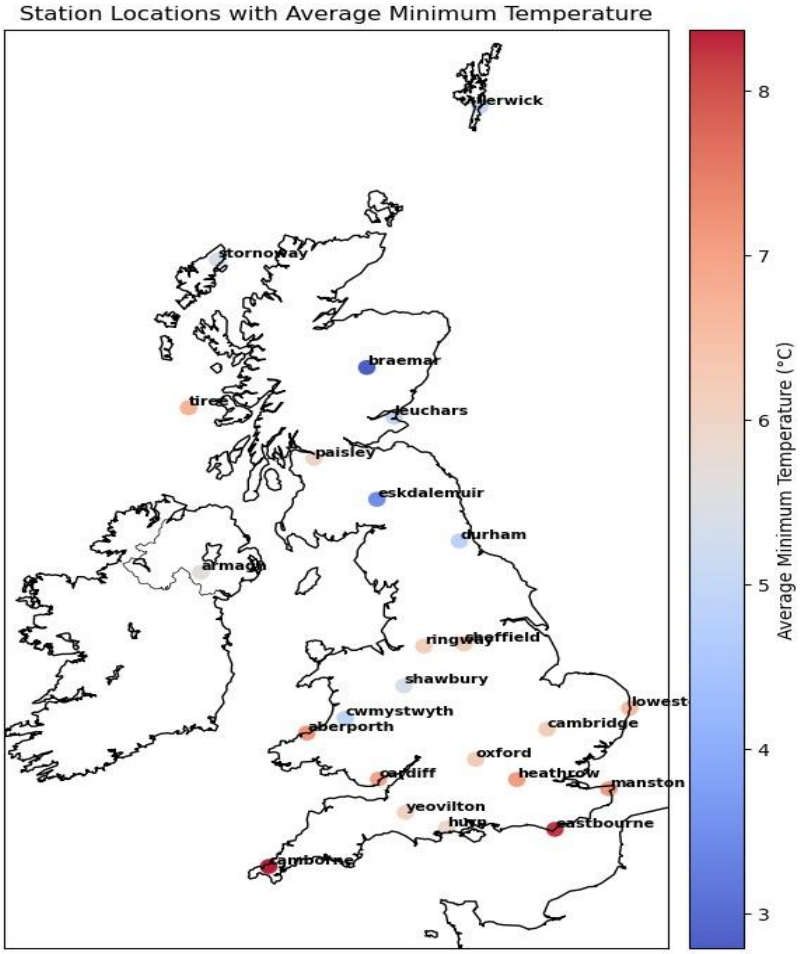


Fig 1 : Map showing stations with average minimum temperature

1. **Geographical Distribution:**
    - The weather stations are spread across the UK, from northern Scotland to the southern coast of England. The map covers diverse environments, including coastal, inland, and elevated regions, reflecting the range of climates across the country.
2. **Temperature Gradients:**
    - **Cooler Regions:** The northern stations, particularly in Scotland (e.g., Braemar, Eskdale Muir, Lerwick), are shaded in dark blue, showing lower average minimum temperatures around 3°C to 4°C. These areas are known for their colder climates, which is visually emphasized in the map.

- **Warmer Regions:** Southern and coastal areas, such as Eastbourne, Heathrow, and Camborne, are shaded in red tones, indicating higher average minimum temperatures closer to 7°C to 8°C. These regions experience milder winters

Chart visualizing average maximum temperatures.

The chart visualizes the average maximum temperatures across weather stations in the United Kingdom, using a geographical map marked with a colour gradient scale ranging from blue (cooler temperatures) to red (warmer temperatures). The data, presented in degrees Celsius (°C), captures the temperature distribution across the UK.

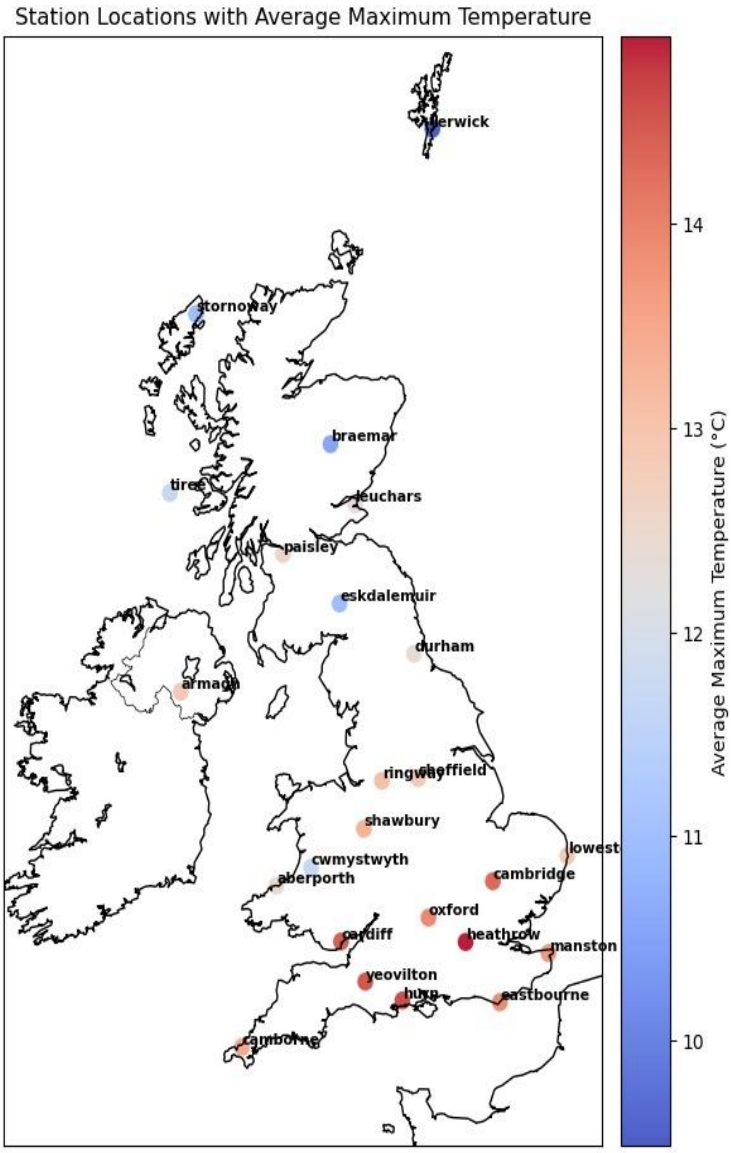


Fig 2 : Map showing locations with average minimum temperature

Observations:

1. **Geographical Spread:**
   - The chart encompasses stations from the northernmost parts of Scotland to the southern coast of England, including Northern Ireland and Wales.
   - The stations are distributed across varying latitudes and altitudes, which likely contribute to the observed temperature variations.
2. **Temperature Distribution:**
   - **Northern Regions:** Stations in Scotland, such as Lerwick, Stornoway, Tiree, and Braemar, are marked in blue and light blue, indicating cooler average maximum temperatures.
   - **Central and Southern England:** Stations like Heathrow, Oxford, and Cambridge are represented in red and orange, showing warmer temperatures. The highest temperatures are concentrated in the southeastern regions, particularly around Heathrow and Oxford.

3 **Regional Temperature Patterns:**

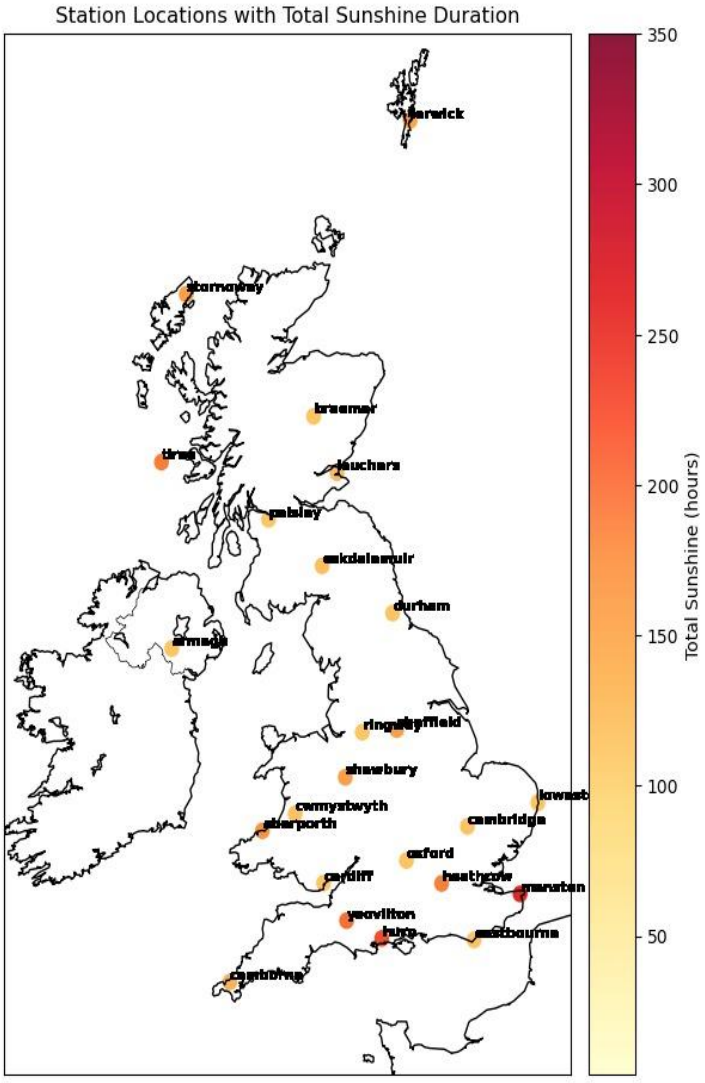
b. **Scotland and Northern Ireland:** Predominantly cooler temperatures are seen in these regions, with blue hues dominating the map, especially at stations like Lerwick and Eskdale Muir.

c. **Wales:** Exhibits moderate temperatures, with stations like Aberporth and Cardiff leaning towards the warmer end of the spectrum.

d. **England:** A clear temperature gradient is visible, with cooler conditions in the north and progressively warmer temperatures in the south. The southeast, especially around Heathrow and Cambridge, records the highest average maximum temperatures.

Chart visualizing locations with total sunshine duration.

The map visualizes total sunshine duration across weather stations in the United Kingdom.



Geographical Coverage:

- The map spans the entire UK, from northern Scotland to southern England, offering a broad perspective on sunshine duration across diverse regions.

Sunshine Gradients:

- Stations like Lerwick in the far north and Manston in the southeast are shown in deep red hues, indicating higher sunshine hours,
- Locations like Camborne and some parts of western Scotland are depicted in lighter yellow shades, indicating lower sunshine hours, typically below 150 hours.
- **Coastal vs. Inland Trends:**
    - Coastal regions, especially in the south such as Eastbourne and Manston, tend to experience higher sunshine duration. These areas benefit from fewer obstacles like mountains and the moderating influence of the sea, which reduces cloud cover.
    - Inland stations, particularly in central and northern regions like Eskdale Muir and Shawbury, show lower sunshine durations, reflecting the cloudier and wetter conditions typical of these areas.

Fig 3 : Map showing stations with total sunshine duration

Cambridge appears to enjoy more sunshine hours than locations further north or west in the UK. This positioning places it among the sunnier areas of the country, at least according to the data visualized in this map. The colour coding suggests Cambridge likely receives between 150-200 hours of total sunshine over the measured period, though the exact figure isn't specified without more detailed scale information.

Notably, Cambridge's location on the map situates it east of Oxford and north of London, both major cities in England. This southeastern positioning contributes to its favourable sunshine statistics, as the southeast of England generally experiences more sunshine than other parts of the UK due to its proximity to continental Europe and distance from the wetter western regions.

Graph representing top 10 stations by average maximum temperature.

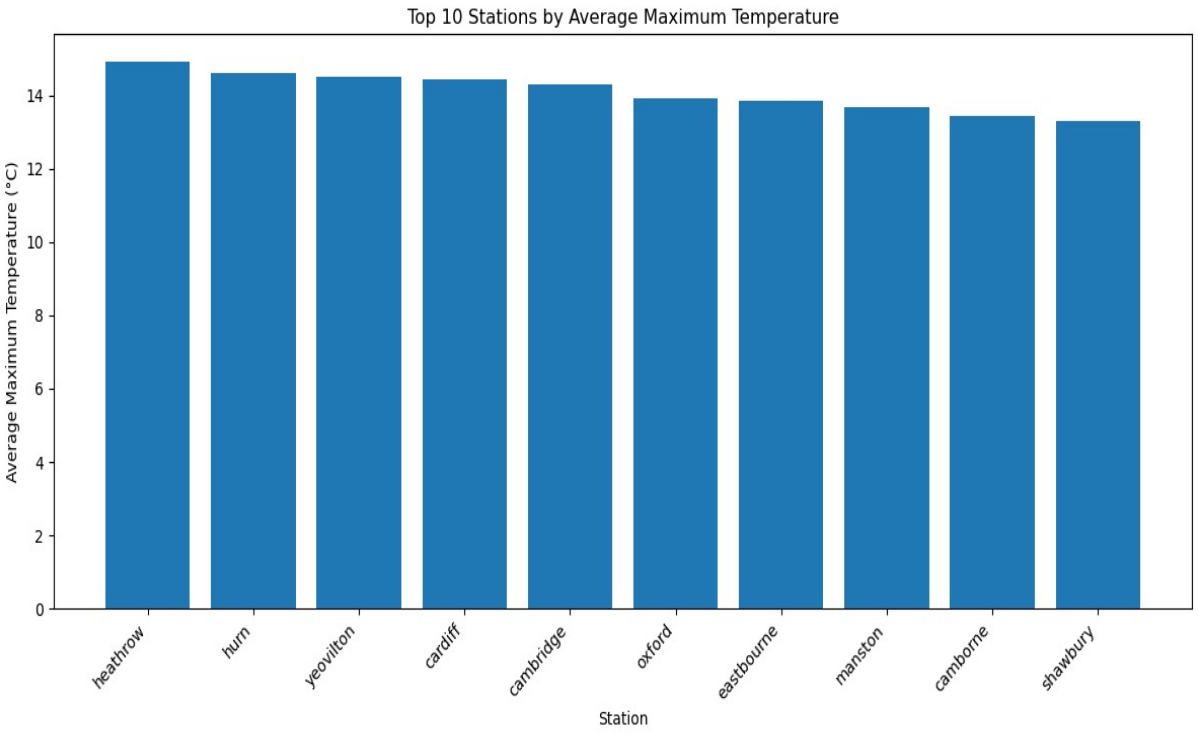


Fig 4 : Bar graph showing top 10 stations by average maximum temperature

The chart highlights the top 10 weather stations ranked by their average maximum temperature in degrees Celsius (°C). Each bar in the chart represents a different weather station, with the bar's height reflecting the average maximum temperature at that location. The stations are displayed along the x-axis, while the y-axis tracks the average temperatures, making it easy to compare the values visually.

Heathrow stands out with the highest average maximum temperature, slightly above 14°C. Following closely are stations like Hurn, Yeovilton, and Cardiff, all showing average maximum temperatures a bit over 14°C.

Graph representing top 10 stations by average sunshine duration

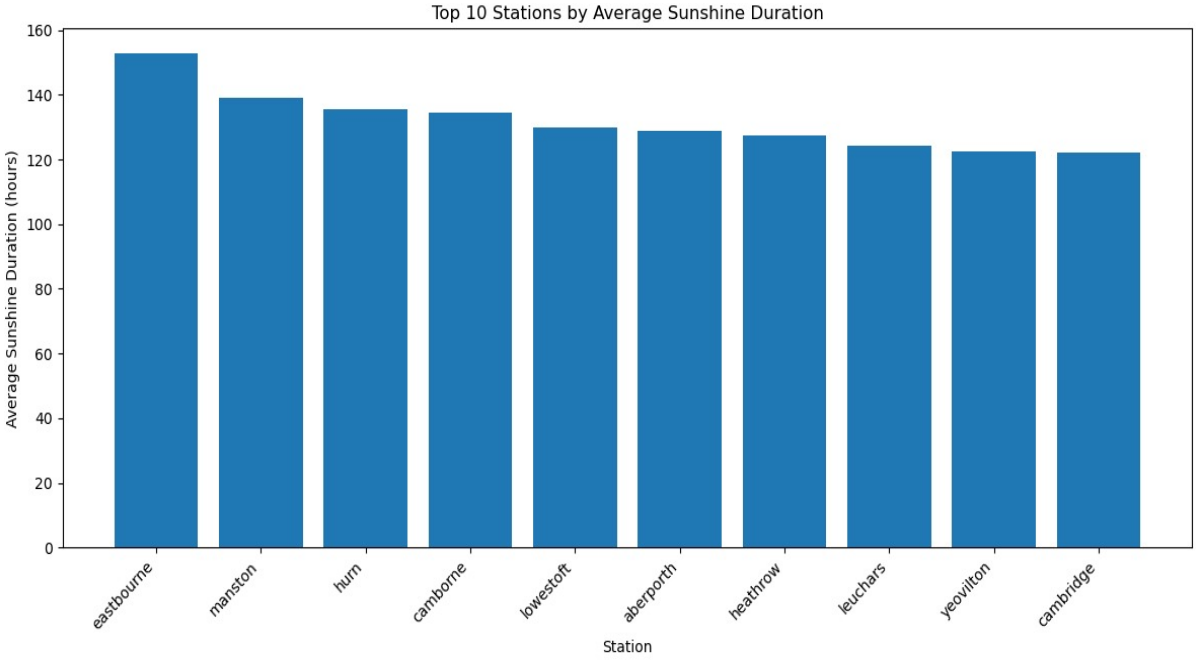


Fig 5: Bar graph showing top 10 stations by average sunshine duration

The chart showcases the top 10 weather stations ranked by their average sunshine duration, measured in hours. Each bar represents a different station, with its height indicating the average sunshine hours recorded at that location. The stations are listed along the x-axis, while the y-axis tracks the average sunshine duration, making it easy to compare the data.

Eastbourne leads the chart with the highest average sunshine duration, approaching 160 hours. Marston follows closely with just under 150 hours of sunshine. Stations like Hurn, Camborne, and Lowestoft also show notable sunshine levels, each averaging between 140 to 145 hours. The remaining stations—Aberporth, Heathrow, Leuchars, Yeovilton, and Cambridge—record similar sunshine durations, ranging from 135 to 140 hours.

In our report the analysis is based on Cambridge and the land surrounding it. Cambridge, UK, can be considered a potentially optimal location for solar panel installation due to several factors:

1. Sunshine Duration: As evidenced by the map, Cambridge receives a relatively high amount of sunshine compared to many other UK locations. This increased solar radiation can lead to higher energy production from solar panels
2. Climate: Cambridge has a temperate maritime climate, with mild temperatures year-round. This is beneficial for solar panel efficiency, as extreme heat can decrease panel performance
3. Land Availability: While Cambridge is a urban area, it has a mix of residential, commercial, and rural surroundings. This diversity provides various opportunities for solar installations, from rooftop systems to potential solar farms in nearby areas (Wiginton et al., 2010).

Satellite Imagery

The satellite images are from the sentinel 2 satellite which are part of the European Union's Copernicus program. The data is acquired from the Copernicus browser where we can select a region of interest and download Raster files for that region and the regions around it.

From the data source dropdown menu, we must make sure that sentinel 2 is selected. We can also choose from what type of data is required, Level-1C (raw data), Level-2A (surface reflectance) and MSI (Multi Spectral Instrument). In our project we have chosen the Multi Spectral Instrument data.

MSI (Multi Spectral Instrument) is the main sensor onboard the Sentinel-2 satellites, The MSI captures the data in 13 different spectral bands ranging from the visible to the shortwave infrared part of the electromagnetic spectrum. These bands allow for a wide range of applications, including land cover classification, vegetation monitoring etc.

The MSI provides imagery at different spatial resolutions:

- 10 meters for visible and near infrared bands
- 20 meters for red edge and shortwave infrared bands

Next, by using the cloud coverage filter we can limit the search to images with low cloud coverage ensuring clear images which is essential for our project so that we can focus on land classification more effectively.

After downloading, the Sentinel 2 imagery will be stored as different bands in JP2 file format, we have use QGIS for further processing of these images.

Digital Elevation Model

The Digital Elevation Model (DEM) was developed using QGIS (Quantum Geographic Information System) which is an open-source geographic information system (GIS) software that allows users to create, visualize, analyze and interpret spatial and geographic data.

To download a digital elevation model using QGIS an additional plugin named SRTM (Shutter Radar Topography mission) downloader is required.

The Shuttle Radar Topography Mission (SRTM), conducted by NASA in February 2000, aimed to create a comprehensive global digital elevation model by using radar to measure Earth's surface elevation. The mission employed synthetic aperture radar (SAR) aboard the Space Shuttle Endeavour to capture detailed topographic data across nearly the entire globe, providing high-resolution elevation maps that are widely used in environmental studies, urban planning, and disaster management. The resulting SRTM data, available at various resolutions, has become a crucial resource for analyzing Earth's terrain.

The SRTM downloader plugin can be easily installed in QGIS by going to the plugins option. After the download we can use the plugin and specify the boundaries (North, West, East, South) of the Region of interest and download the DEM. Before downloading it is essential to register and create a free ID at https://urs.earthdata.nasa.gov/ which is NASA's Earth data system, which is part of the Earth Science Data and Information System (ESDIS) Project.

After successfully downloading the DEM, we can use it to calculate slope for the region of interest. The downloaded DEM will not necessarily be the same size as the satellite imagery and resampling of the images is required, which is later done in our project.

Analytical Techniques (Himanshu Sharma)

Sentinel – 2 Spectral bands

To classify between land types such as vegetation, urban areas, water bodies we have used multiple spectral bands in our project.

The bands used in our project are:

1. **Blue Band (Band 2)**:
    - **Path**: T30UYC_20240126T111331_B02_10m.jp2
    - **Resolution**: 10 meters
    - **Usage**: The blue band is useful for coastal and water body monitoring, as well as for analyzing vegetation and urban areas. It helps in distinguishing between different types of land cover.
2. **Green Band (Band 3)**:
    - **Path**: T30UYC_20240126T111331_B03_10m.jp2
    - **Resolution**: 10 meters
    - **Usage**: The green band is critical for vegetation analysis and detecting plant health. It's also used for assessing water quality and vegetation cover.
3. **Red Band (Band 4)**:
    - **Path**: T30UYC_20240126T111331_B04_10m.jp2
    - **Resolution**: 10 meters
    - **Usage**: The red band is essential for vegetation analysis, including calculating vegetation indices like NDVI (Normalized Difference Vegetation Index), which helps in assessing plant health and land cover.
4. **Near-Infrared Band (Band 8)**:
    - **Path**: T30UYC_20240126T111331_B08_10m.jp2
    - **Resolution**: 10 meters
    - **Usage**: The near-infrared band is crucial for vegetation health analysis and land cover classification. It is used in indices like NDVI to evaluate plant vigour and to distinguish between different types of vegetation.
5. **Short-Wave Infrared Band (Band 11)**:
    - **Path**: T30UYC_20240126T111331_B11_20m.jp2
    - **Resolution**: 20 meters
    - **Usage**: The short-wave infrared band is useful for detecting moisture content in soil and vegetation, as well as for assessing different types of minerals and geological features. This band is effective for distinguishing urban areas due to its sensitivity to moisture and surface materials. Urban surfaces such as roads

and buildings have distinct reflectance characteristics in this band. It is useful for urban heat island studies and assessing surface composition.

Reading the bands

To read the sentinel 2 bands in our code we use the rasterio library which is a Python library designed for reading and writing raster data, which is commonly used in geographic information systems (GIS) and remote sensing.

We create a function using rasterio which reads a JP2 file to extract the image data, the affine transformation used to map the image to geographic coordinates, and the coordinate reference system (CRS) of the image. This information is crucial for analysing and interpreting geospatial data..

Stacking the bands and extracting ROI

After reading the raster bands we stack the bands to create a satellite image, before doing so we extract our Region of interest from the image which is Cambridge. To do so we have to know the pixel coordinates of the ROI i.e. the values x_min, x_max, y_min, y_max which then can be entered in the function which will calculate the region according to the pixel coordinates, stack the bands and generate the required satellite image.

This approach not only focuses on extracting a region of interest but also has several other benefits:

- **Faster Processing**: By selecting a smaller region of interest, the amount of data that needs to be processed is reduced, which can significantly speed up computations and analyses. This is especially useful when working with high-resolution or large-scale datasets where processing the entire dataset might be inefficient or impractical.
- **Focused Analysis**: Extracting a specific ROI allows for more detailed and focused analysis within a particular area of interest. This is often done to study specific features, assess changes, or conduct detailed spatial analysis.
- **Memory Efficiency**: Working with a smaller subset of data reduces memory usage, making it easier to handle within limited computational resources. This is crucial for avoiding memory overflow and ensuring that processing can be completed efficiently.

Overall, we are extracting a defined rectangular region from various raster datasets (including multiple spectral bands) to limit the analysis to a specific area. This approach optimizes

processing time and memory usage by focusing on a smaller, more manageable portion of the data, making it easier to perform detailed analyses and visualizations within the selected region.

NDVI (Normalized Difference Vegetation Index)

Normalized Difference Vegetation Index is a widely used remote sensing measurement that assesses the health and density of vegetation on the Earth's surface. It utilizes satellite or aerial imagery to calculate the difference between the near-infrared (which vegetation strongly reflects) and red light (which vegetation absorbs) to create a ratio. This ratio is standardized, or normalized, to produce an index value that ranges from -1 to +1. In practical terms, healthy, dense vegetation typically shows high NDVI values close to +1, indicating robust plant health and coverage. Conversely, areas with sparse or stressed vegetation exhibit lower NDVI values.

NDVI and its relevance to solar panel installation:

1. **How NDVI Works**:
    - NDVI leverages the fact that healthy vegetation strongly reflects **near-infrared (NIR)** light due to chlorophyll content, while absorbing **red** light for photosynthesis.
    - By comparing NIR and red reflectance, NDVI highlights variations in vegetation cover across a landscape.
2. **Applications of NDVI**:
    - **Vegetation Monitoring**: NDVI is widely used to assess vegetation health, monitor crop growth, detect stress, and estimate biomass.
    - **Land Use and Land Cover Studies**: NDVI helps classify land cover types (e.g., forests, croplands, grasslands) and track changes over time.
    - **Environmental Studies**: NDVI informs ecological research, including habitat assessment and conservation efforts.

3. **NDVI in Solar Panel Installation**:
    - **Site Selection**: NDVI maps can guide solar panel placement by identifying areas with minimal vegetation cover. Open spaces with low NDVI values are ideal for solar installations.
    - **Shading Analysis**: NDVI data can assess potential shading from nearby trees or structures. Panels should avoid shaded areas to maximize energy production.

- **Maintenance Planning**: Regular NDVI monitoring helps detect vegetation encroachment near solar panels. Timely trimming prevents shading and maintains panel efficiency.

Calculating NDVI

NDVI is calculated using the red and near-infrared (NIR) bands of the electromagnetic spectrum. The formula is:

$$\text{NDVI} = \frac{NIR - RED}{NIR + RED}$$

Where:

- **NIR**: Reflectance in the near-infrared band.
- **Red**: Reflectance in the red band.

In our code, First, an array named ndvi is initialized with zeros. This array is of the same shape as the NIR band, ensuring that each pixel in the NDVI output will have a corresponding value. The array is set to type float to accommodate the decimal values resulting from the NDVI calculation.

Next, to avoid potential errors during division, a mask is created. This mask identifies pixels where the sum of the NIR and Red band values is not zero. If the sum of these two bands is zero, division by zero would occur, which could result in undefined or erroneous NDVI values. The mask is a Boolean array where each position is True if the sum of the NIR and Red values is non-zero, and False otherwise.

With the mask in place, the NDVI is then calculated only for the pixels where the mask is True. This means the calculation is performed only on those pixels where both NIR and Red values are valid, thus avoiding the division by zero issue. The NDVI formula is applied to these selected pixels: it subtracts the Red band value from the NIR band value and then divides by the sum of the two bands.

$$\text{NDVI} = \frac{(band\_nir[mask].astype(float) - band\_red[mask].astype(float))}{(band\_nir[mask] + band\_red[mask])}$$

The resulting NDVI values are then stored in the ndvi array at the corresponding positions.

In summary, the process ensures that NDVI values are accurately computed by initializing an array to store results, using a mask to avoid division errors, and applying the NDVI formula only where valid data is available. This approach helps in producing a reliable NDVI output that can be used for analysing vegetation health and density in remote sensing applications.

In our code we have displayed the resulting NDVI

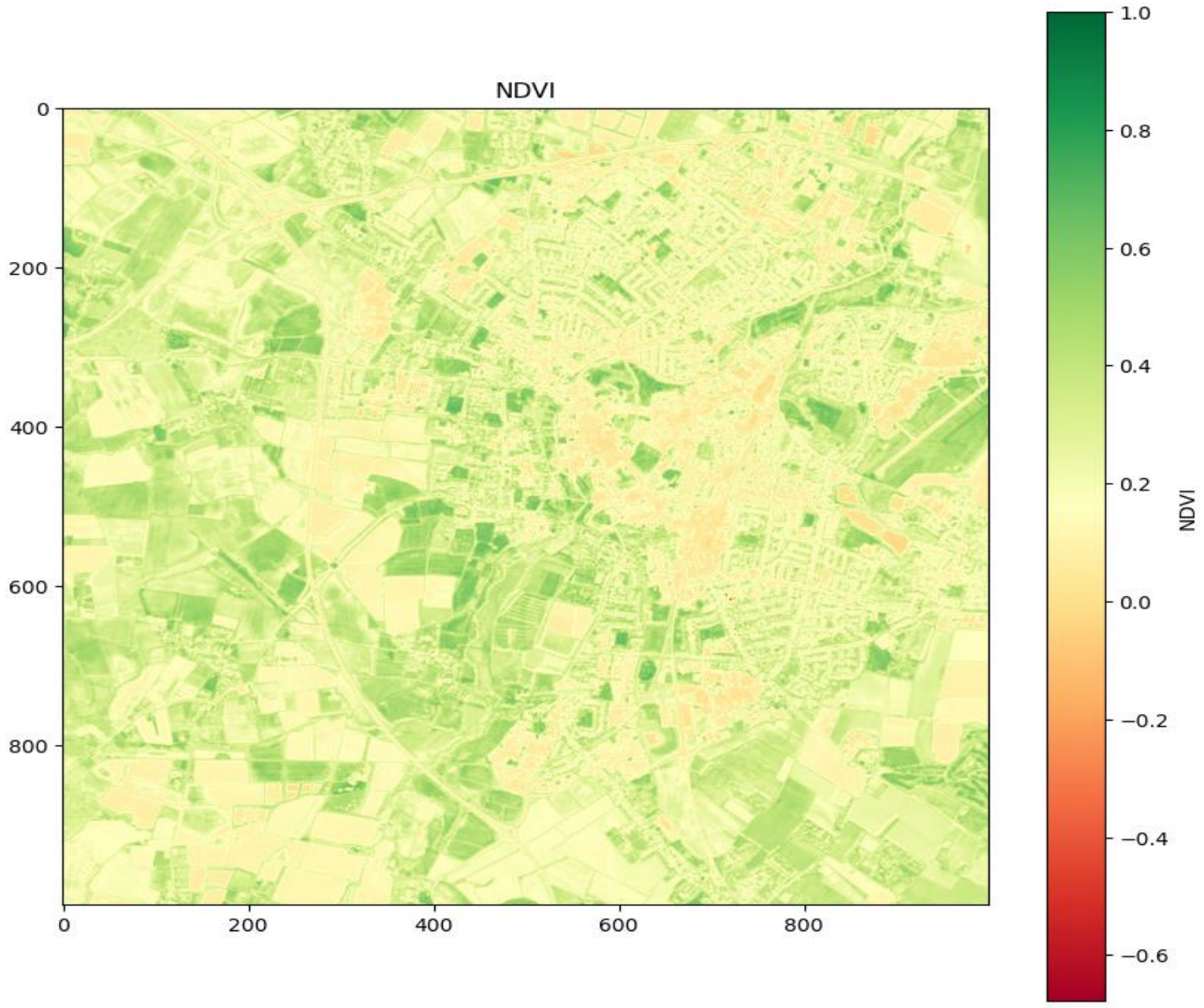


Fig 6: Map displaying NDVI values across Cambridge

The area's highlighted in green have good vegetation health whereas the areas with brown or light brown shade do not have or have a little bit of vegetation.

NDBI (Normalized Difference Built-up Index)

The **Normalized Difference Built-up Index (NDBI)** is a remote sensing index used to detect built-up or urbanized areas based on the variation in surface reflectance between **near-infrared (NIR)** and **shortwave infrared (SWIR)** bands. The Normalized Difference Built-up Index (NDBI) is a remote sensing metric used to identify and analyze built-up or urban areas in satellite and aerial imagery. Calculated using the Short-Wave Infrared (SWIR) and Near-

Infrared (NIR) bands, NDBI highlights areas with high SWIR reflectance relative to NIR reflectance, which is typical of urban surfaces like concrete and asphalt. Positive NDBI values generally indicate built-up areas, while negative values are associated with non-built-up regions such as vegetation or water.

1. **How NDBI Works**:
    - Urban areas typically have high reflectance in the SWIR band due to the presence of man-made structures (concrete, asphalt, etc.).
    - Vegetation and natural surfaces have lower SWIR reflectance, resulting in negative NDBI values.
2. **Applications of NDBI**:
    - **Urban Mapping**: NDBI helps identify built-up regions, urban sprawl, and land-use changes.
    - **Heat Island Effect**: NDBI indirectly reflects the urban heat island effect, where built-up areas retain more heat than natural landscapes.
3. **Solar Panel Installation**:
    - **Site Selection**: NDBI maps guide solar panel placement by identifying suitable built-up areas with minimal vegetation cover.
    - **Shading Analysis**: NDBI assists in assessing potential shading from nearby buildings or infrastructure.

Calculating NDBI

NDBI is calculated using the short-wave infrared and near infrared bands of the electromagnetic spectrum. The formula is:

$$\text{NDBI} = \frac{SWIR - NIR}{SWIR + NIR}$$

Where:

- **(SWIR) represents the reflectance in the shortwave infrared band.**
- **(NIR) represents the near-infrared reflectance.**

In our code, The Normalized Difference Built-up Index (NDBI) by initializing an array to store the results, creating a mask to prevent division by zero errors, and then applying the NDBI formula only to valid pixels. It first initializes an array ndbi of the same size as the Near-Infrared (NIR) band with zeros. A boolean mask is created to identify pixels where the sum of the Short-

Wave Infrared (SWIR) and NIR bands is non-zero, ensuring that calculations are performed only where valid data is present.

$$\text{NDBI[mask]} = \frac{(resampled\_swir[mask].astype(float) - band\_nir[mask].astype(float))}{SWIR(resampled\_swir[mask] + band\_nir[mask]) + NIR}$$

The NDBI is computed by subtracting the NIR reflectance from the SWIR reflectance, dividing by their sum, and storing the result in the ndbi array at the corresponding pixel locations. This method ensures accurate and error-free NDBI calculations for identifying built-up areas.

Land Cover Classification

After calculating the NDVI and NDBI values for the satellite images we can move forward with the land cover classification. To move forward with the land cover classification, we assign threshold values to characteristics such as land vegetation water etc of the image. This is done as follows:

Initially, an array called labels is created, with the same number of elements as there are pixels in the raster images. This array is initialized to zeros, which will serve as the default label for all pixels. Each label corresponds to a different land cover type.

The first condition labels pixels as vegetation if their NDVI value exceeds 0.3. NDVI values greater than this threshold generally indicates healthy vegetation, so these pixels are assigned the label 1 to represent vegetation.

The second condition assigns a label of 2 for water bodies. This is determined by checking if both the Near-Infrared (NIR) band value is less than 1000 and the red band value is less than 500. These thresholds are used because water bodies typically have low reflectance in both the NIR and Red bands.

We split the dataset into training and testing subsets, then we train a Random Forest classifier using the training data. Initially, the ‘train_test_split’ function from the ‘sklearn’ library is used to divide the combined feature matrix data and the associated labels into two subsets: a training set and a testing set. Here, 30% of the data is reserved for testing, while the remaining 70% is used for training the model. This split is done to evaluate the performance of the classifier on unseen data, ensuring that it generalizes well.

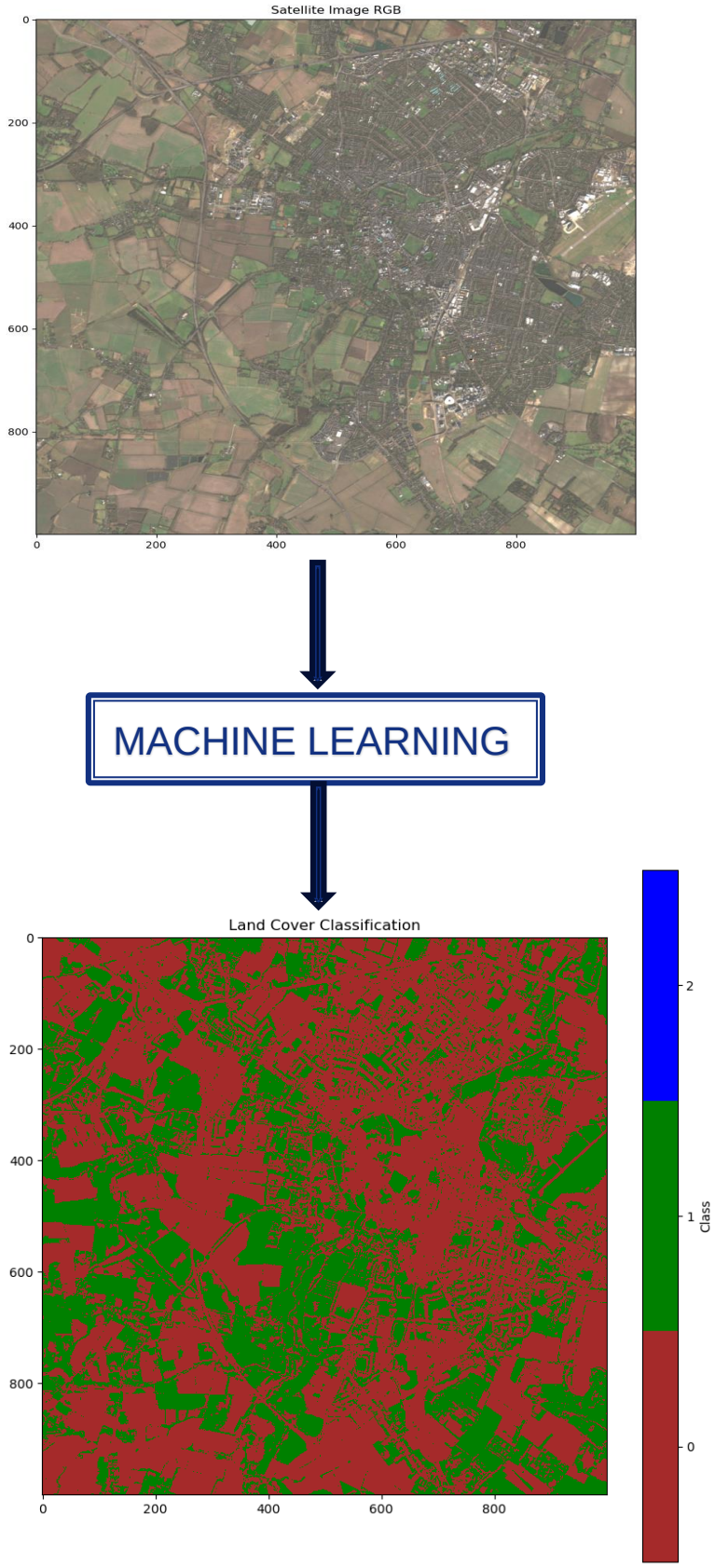


Fig 7: Conversion of map of Cambridge to one showing its Land cover classification.

The fit method is used to train the Random Forest model on the training data. During this phase, the model learns the relationships between the features (such as the spectral bands and indices) and the labels (land cover types). The decision trees within the forest analyze various aspects of the data to classify pixels into categories such as vegetation, water, or urban areas. The learning process involves finding patterns and making predictions based on the input features.\

Land cover classification is essential for optimizing solar panel installations. It helps in selecting suitable sites by identifying areas with minimal shading and avoiding obstacles. By tracking changes in land cover, it aids in assessing the impact of new solar projects on existing land use. It also supports environmental impact evaluations, infrastructure planning, and compliance with regulations. Ultimately, it ensures that solar panels are placed effectively to maximize energy production and minimize environmental disruption.

Land cover classification plays a crucial role in optimizing solar panel installations. By accurately identifying and categorizing different types of land cover.

Otsu’s Thresholding Method

Otsu’s method is a widely used image thresholding technique that automatically determines the optimal threshold value to separate the foreground from the background in an image. It works by minimizing the intra-class variance, which is the variance within each of the two classes (foreground and background) created by the threshold.

In land cover classification, Otsu’s method can be particularly useful for distinguishing between different types of land cover, such as built-up areas, vegetation, and water bodies. This method helps in accurately mapping and monitoring urbanization and other land cover changes by providing a clear distinction between different land cover types.

In our code, we have used Otsu's thresholds for NDVI, NDBI, NIR, and SWIR by flattening each of these indices and spectral bands into a one-dimensional array. Otsu's method is then applied to these arrays to determine the optimal threshold values that best separate the pixel values into two classes: typically, one representing the target land cover type and the other representing the rest of the pixels.

With these thresholds determined, the code proceeds to classify the pixels in the image into different land cover types. First, it initializes an array called labels with zeros, which will later store the classification results for each pixel. Using the thresholds derived from Otsu's method, the code assigns a label of 1 to pixels that are classified as vegetation (where NDVI exceeds the threshold). Water pixels are identified by either a high NDVI (beyond a certain adjusted threshold) or a combination of low NIR and Red band values, and these pixels are labeled as 2. Urban areas are identified by a combination of high SWIR and low NDVI values, and these pixels are labeled as 3. Finally, any remaining pixels that do not fall into the previous categories are labeled as 4, representing general land cover.

The labeled pixels are then reshaped back into the original image's 2D shape to match the spatial structure of the input image. For visualization purposes, a color map is defined where different colors represent different land cover types: green for vegetation, blue for water, grey for urban areas, and brown for land. This allows for easy interpretation and analysis of the classified land cover types based on the spectral characteristics captured in the remote sensing imagery.

Application of Otsu's Threshold in Land Cover Classification:

When there is no ground truth data available for land cover classification, Otsu's thresholding can be a useful tool for unsupervised classification of satellite or aerial images. Here’s how it can be applied:

1. **Binary Classification**: Otsu's method is particularly effective for binary classification problems, such as distinguishing between water and land, or vegetation and non-vegetation areas.

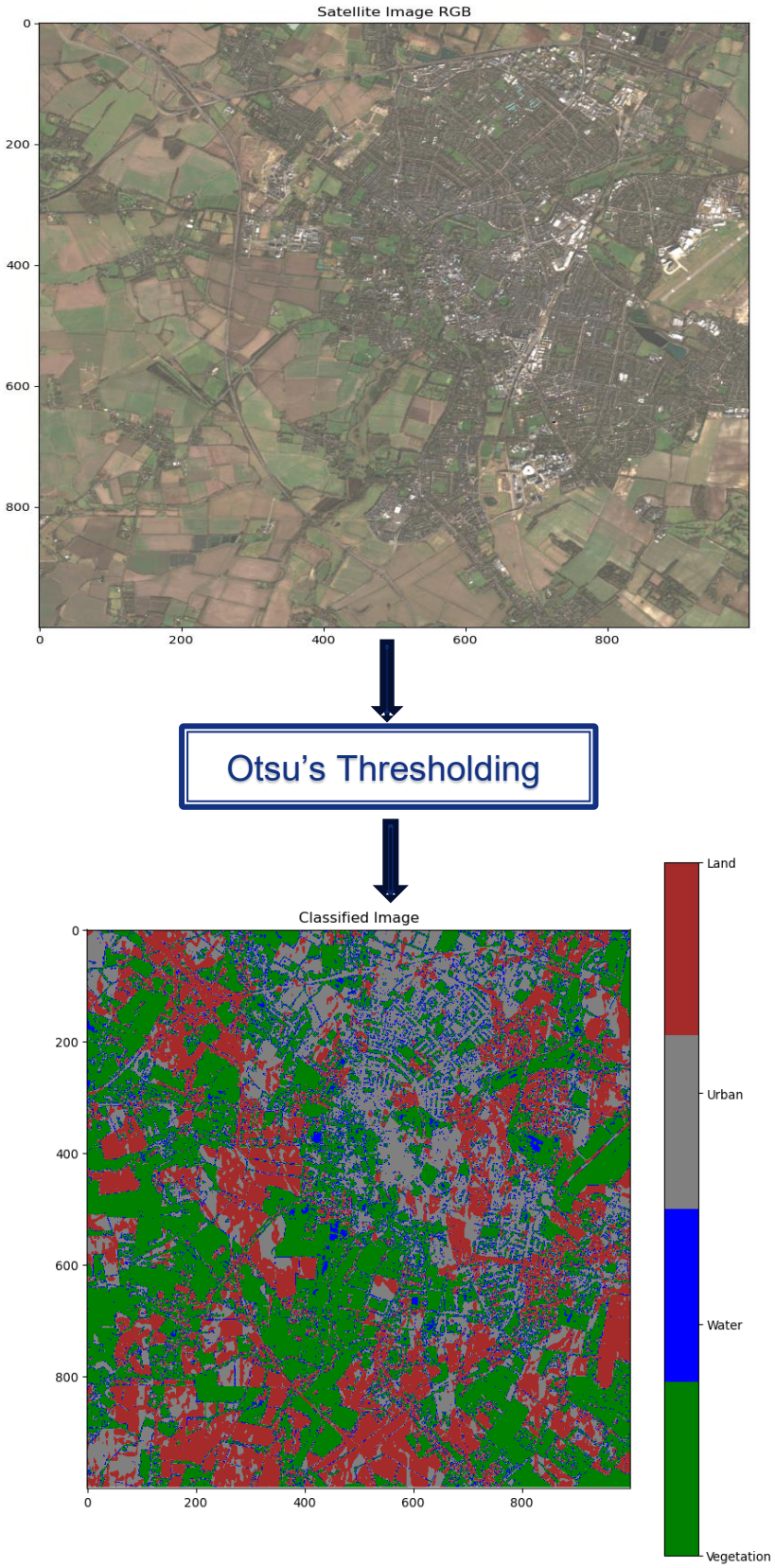


Fig 8 : Land cover classification by Otsu's Thresholding method

2. **Thresholding on Spectral Indices**: You can apply Otsu's thresholding to specific spectral bands or indices (like NDVI, NDBI, or NDWI) derived from satellite imagery. This can help segment the image into two primary classes (e.g., urban vs. non-urban, water vs. land) without requiring ground truth data.
3. **Unsupervised Land Cover Segmentation**: Otsu's method can also be extended to multiple classes by iteratively applying it to different segments of the image or by using multi-threshold techniques. This approach allows for the unsupervised segmentation of land cover types, which can then be interpreted based on their spectral properties.
4. **Preprocessing for Further Classification**: Otsu's thresholding can serve as a preprocessing step in more complex unsupervised classification algorithms, helping to roughly segment the image into areas of interest before applying more detailed analysis.

Setting thresholds using Otsu's method

In the code, thresholds for different land cover types are being set using Otsu's method, which optimizes the separation of pixel values into different categories by maximizing the variance between these categories. Here's how the thresholds are applied to classify the land cover:

1. Vegetation Classification:

- **Condition:** NDVI > Otsu's NDVI Threshold
- **Explanation:** The NDVI (Normalized Difference Vegetation Index) is a common index used to identify vegetation. Pixels with NDVI values greater than the threshold calculated by Otsu's method (around 0.2694) are likely to represent areas with healthy vegetation. Thus, any pixel with an NDVI value above this threshold is labelled as vegetation (labels = 1).

2. Water Classification:

- **Condition:** NDVI > (Otsu's NDVI Threshold + 0.3) **or** (NIR < Otsu's NIR Threshold and Red < (Otsu's NIR Threshold / 3))
- **Explanation:** Water bodies generally have low NDVI values because water absorbs most of the NIR radiation and reflects less in the red spectrum. The first part of the condition NDVI > (Otsu's NDVI Threshold + 0.3) ensures that only pixels with significantly higher NDVI values are considered water, reflecting the distinct spectral signature of water. The second part of the condition identifies water by checking if the NIR band values are lower than Otsu's threshold (3087), and if the red band values are also low, confirming the pixel as water (labels = 2).

3. Urban Area Classification:

- **Condition:** SWIR > Otsu's SWIR Threshold **and** NDVI < Otsu's NDVI Threshold
- **Explanation:** Urban areas typically exhibit higher SWIR (Short-Wave Infrared) values because of their surfaces (like concrete and asphalt) which reflect SWIR more than other land covers. Simultaneously, these areas often have low NDVI values due to the absence of vegetation. Therefore, pixels that have SWIR values greater than Otsu's SWIR threshold (2661) and NDVI values lower than the NDVI threshold are classified as urban areas (labels = 3).

4. Remaining Land Classification:

- **Condition:** Pixels not classified as vegetation, water, or urban are classified as land.
- **Explanation:** The remaining pixels, which do not meet the criteria for vegetation, water, or urban, are assumed to be land. These could include bare soil, dry areas, or other non-

vegetated land types. This catch-all classification ensures that all pixels in the image are accounted for (labels = 4).

Slope

Calculating slope from DEM

In the code, we calculate slope from the DEM by using a function which is explained as follows:

The process begins by calculating the gradient of the DEM in the x (horizontal) and y (vertical) directions. The gradient represents the rate of elevation change along these two axes. This calculation is essential because it gives an understanding of how steep the terrain is in each direction. The function ‘np.gradient’ is used for this purpose, and it requires the pixel size in both directions, provided by transform[0] and transform[4]. These values are critical because they scale the gradient correctly according to the actual distance on the ground, ensuring that the slope is accurately represented.

Once the gradients in the x and y directions are obtained, the next step is to calculate the slope. The slope is a measure of the steepness or incline of the terrain at each point.

Understanding the DEM and it’s role

A DEM is a grid-based representation of the Earth's surface, where each cell or pixel contains a single value representing the elevation at that location. This elevation data is crucial for various spatial analyses because it allows us to understand the physical geography of an area. By analysing how elevation changes across the grid, we can derive important terrain attributes, such as slope, aspect, and curvature. The slope, in particular, is vital for understanding the incline or steepness of the terrain, which influences water flow, soil erosion, and even human activities like construction or agriculture.

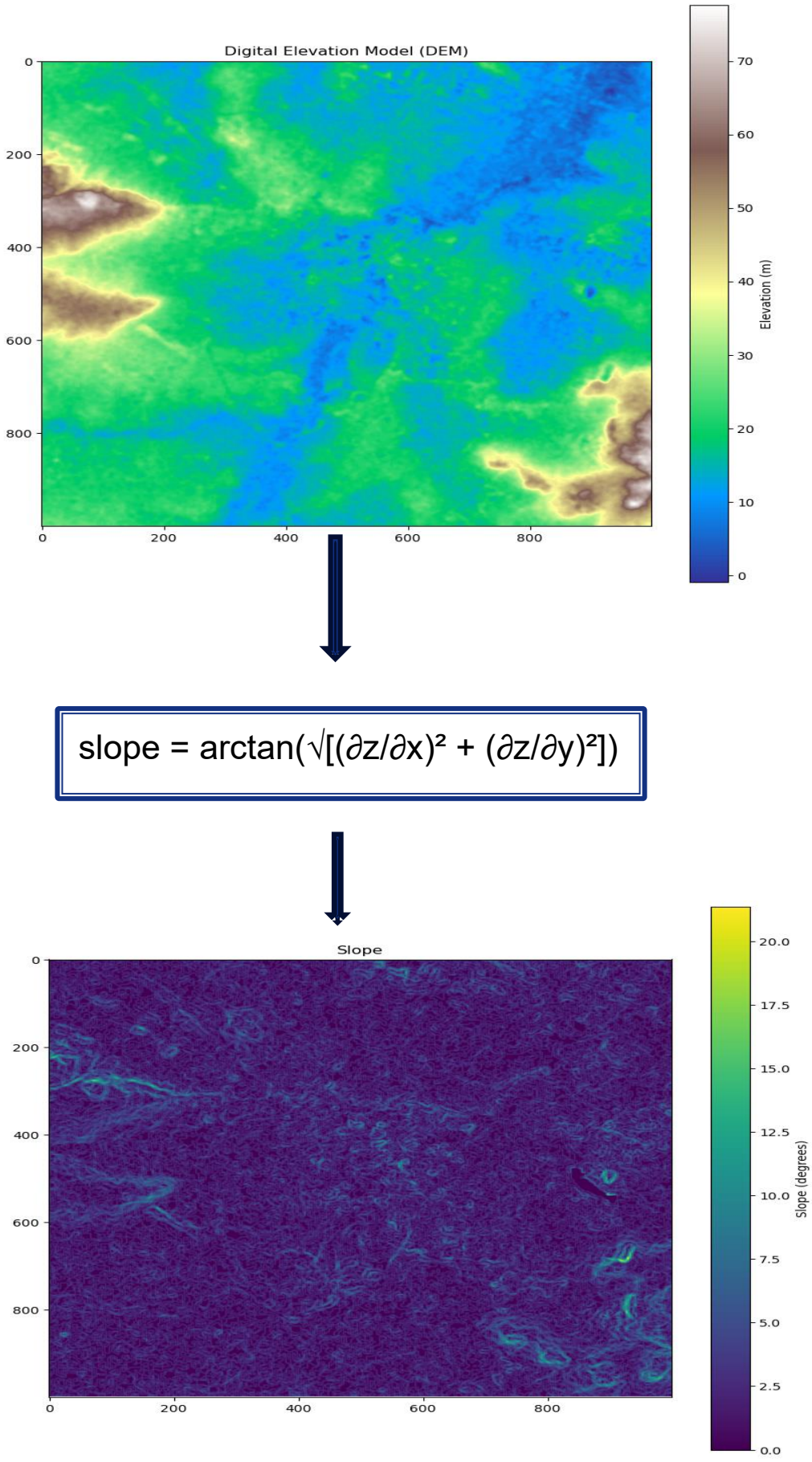


Fig 9 : Conversion of Digital Elevation model to slope map

Mathematical Interpretation of the slope function

The function calculates the gradient of the Digital Elevation Model (DEM) in both x and y directions: ∇DEM = (∂z/∂x, ∂z/∂y) Where:

- ∂z/∂x is the rate of change in the x-direction
- ∂z/∂y is the rate of change in the y-direction

The slope is then calculated using these gradients: slope = arctan(√[(∂z/∂x)² + (∂z/∂y)²])

So, the complete mathematical formula for the slope calculation can be expressed as:

$$\text{slope} = \arctan(\sqrt{[(\partial z/\partial x)^2 + (\partial z/\partial y)^2]})$$

Where:

- z represents the elevation values in the DEM
- x and y represent the horizontal coordinates
- ∂z/∂x and ∂z/∂y are the partial derivatives of z with respect to x and y, respectively

Machine Learning Techniques

Setting Thresholds for Solar Panel Installation

In this approach we have used some thresholds which can be considered as reasonable for determining potential for solar panel locations. We have set NDVI at 0.2 because NDVI at this value and values below this represents non vegetative areas, which are suitable for solar panel installation. A slope threshold is also set at 15 degrees and is considered practical in this case as steeper slopes may reduce solar panel efficiency.

Create a Binary Mask for Potential Solar Panel Locations

A binary mask is created to identify areas suitable for solar panels based on NDVI, slope, and land cover classification. The mask excludes water bodies (class 2) and urban areas (class 3), focusing on areas that are non-vegetative and have a gentle slope.

Feature Preparation for Modelling

- **Features:** Spectral bands (blue, green, red, near-infrared), a resampled short-wave infrared (SWIR) band, NDVI, NDBI (Normalized Difference Built-up Index), elevation data (DEM), and slope are used as features for the models.
- **Labels:** The binary mask created earlier is flattened and used as the target variable (y) for classification.

Data Splitting

- The data is split into training and testing sets using an 70-30 split ratio. This helps in evaluating model performance on unseen data.

Random Forest Regressor with Thresholds

- **Training:** A Random Forest Classifier is trained on the prepared features to classify areas as suitable (1) or not suitable (0) for solar panel installation.
- **Prediction:** Predictions are made on the test set and the entire dataset to visualize suitable locations for solar panels.
- **Visualization:** The predicted suitable areas are visualized on a map, showing binary values indicating suitability.

However, by using the above approach and setting explicit thresholds for NDVI and slope (NDVI < 0.2 and slope < 15 degrees) and use these thresholds to create a binary mask for potential solar panel locations, we are predefining the criteria for suitability. By using this approach, the model will be trained on the data where suitability is already determined by these thresholds, This in turn will limit the models ability to learn these relationships on its own.

To avoid this and allowing the model to learn and generalize better we then avoid setting strict thresholds like before and allow the model to work with raw feature values so it can learn the underlying relationships from the data. This approach will likely lead to a more robust and flexible model that will perform even when applied to new or slightly different data.

Here is how we took the approach:

Random Forest Regressor

- **Target Variable:** A more complex target variable (solar_potential) is created by combining factors like lower NDVI, lower slope, higher elevation, and the exclusion of water and urban areas.
- **Modeling:** A Random Forest Regressor is used to predict a continuous solar potential score, rather than a binary outcome.
- **Evaluation:** The model is evaluated using Mean Squared Error (MSE) and R-squared (R^2) scores. Predictions are visualized on a map.

XGBoost Regressor

- **Training:** The XGBoost Regressor, known for its efficiency and performance, is trained on the features.
- **Evaluation:** The model is evaluated similarly using MSE and R^2, and the results are visualized.
- **Comparison:** A True vs. Predicted plot is created to compare model predictions with actual values.

Gradient Boosting Regressor

- **Training:** A Gradient Boosting Regressor is trained, which iteratively improves predictions by combining weak models.
- **Evaluation and Visualization:** Similar evaluation and visualization steps are performed, as with the Random Forest and XGBoost models.

Artificial Neural Network (ANN)

- **Scaling:** The features are scaled using StandardScaler to ensure they are on a similar scale, which is crucial for neural networks.
- **Model Architecture:** An ANN with three hidden layers is created and trained on the scaled features.
- **Evaluation:** The model's performance is evaluated using MSE and R^2, with a True vs. Predicted plot created for visualization.

Deep Neural Network (DNN)

- **Model Architecture:** A more complex DNN with five hidden layers is created to potentially capture more intricate patterns in the data.
- **Training and Evaluation:** The DNN model is trained, and its performance is evaluated similarly to the ANN. Predictions are visualized on a map, and a True vs. Predicted plot is generated.

Results & Conclusion (Ankit Mate)

Results from the models

We have generated graphs for each of the models used and plotted maps which visualise the predicted solar potential for the Cambridge area using different machine learning models. Each map represents the results from the respective models: Artificial Neural Network (ANN), Deep Neural Network (DNN), Gradient Boosting, and XGBoost.

1. Artificial Neural Network (ANN)

The scatter plot shows the true values on the x-axis and the predicted values on the y-axis. The dashed line represents the ideal case where the predictions exactly match the true values. In this case, most of the points are closely aligned with the line, indicating that the ANN model is making very accurate predictions.

- Mean Squared Error (MSE): 0.000044759805030639436
- R-squared Score (R^2): 0.9969542564503776 which is ~ 0.997

The ANN model has a very low MSE, indicating it makes only minor errors in its predictions. The high R^2 score of 0.997 shows that the model explains 99.7% of the variance in the data, which is an excellent performance.

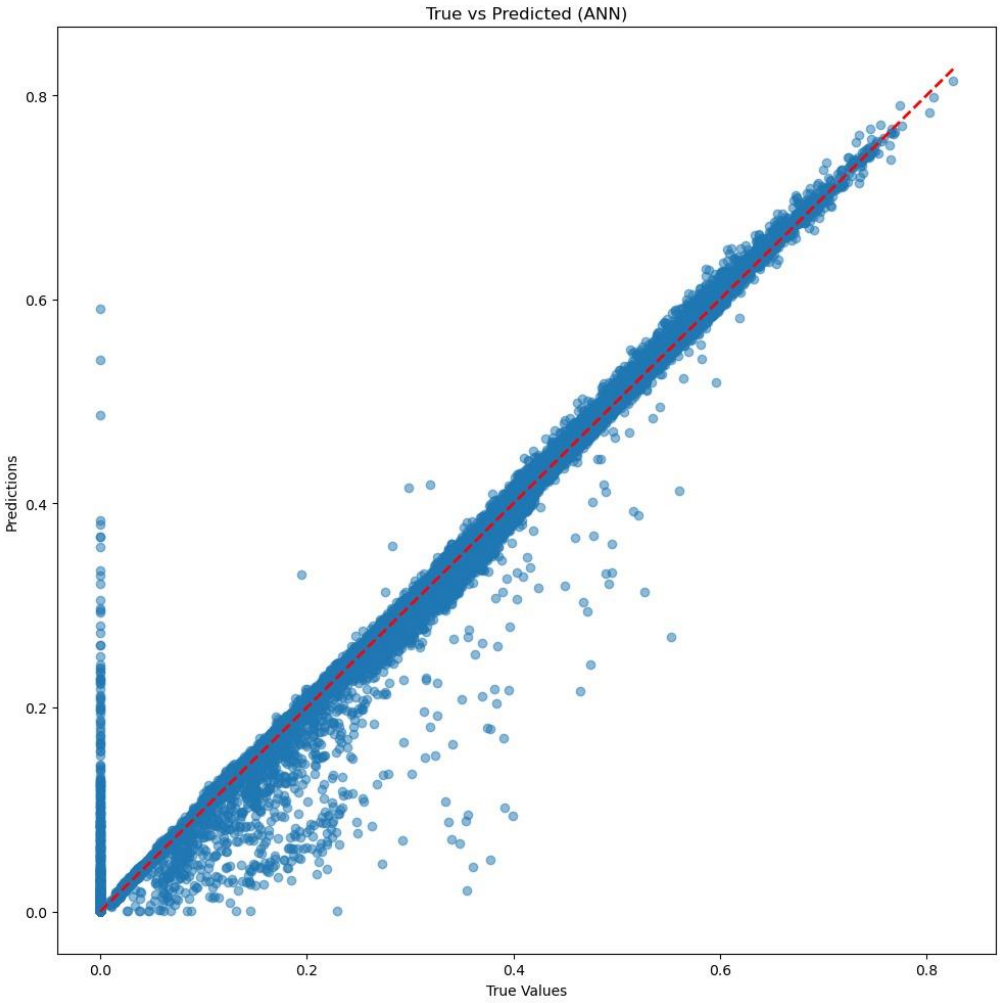


Fig 10: ANN: True V/s Predicted Values

Understanding Model Loss During Training

The graph below helps us understand how well our model is learning from the data during the training process. The key idea is to see if our model is getting better at predicting what we want it to predict.

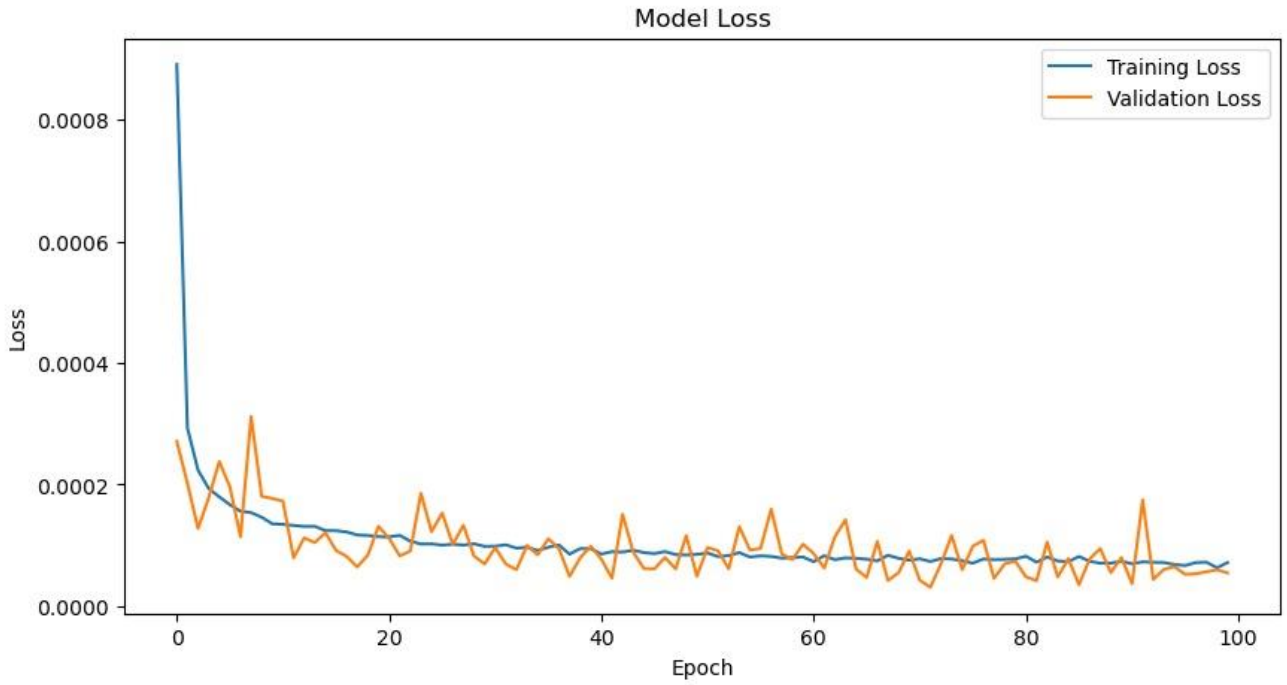


Fig 11 : Chart showing model loss for ANN

Description of the graph

The x-axis represents the number of epochs. An epoch is one complete pass through all the training data. We trained our model for 100 epochs. The y-axis (vertical) represents the loss. Loss is a measure of how wrong the model's predictions are compared to the actual outcomes. Lower loss means better performance.

Training Loss Line (Blue): This shows how well the model is doing on the training data, the data it has already seen and learned from.

Validation Loss (Orange): This shows how well the model is doing on new data (the validation data). This is important to see if the model can generalise well to unseen data.

Understanding the graph

1. Initial Training Phase (First 10 Epochs):

Steep Drop in Loss: Both the training loss and validation loss start high and drop quickly. This is because the model is initially making a lot of errors but quickly starts to learn and improve.

Learning Basics: In these early epochs, the model is learning the basic patterns in the data, which is why we see such a significant improvement.

2. Middle Phase (Around 20-50 Epochs):

Loss Stabilises: After the initial drop, the losses start to become level. The model is still learning, but the improvements are smaller because it is fine-tuning its understanding of the data.

Training and Validation Loss Close Together: Both lines are close together, which is a good sign. It means the model is not just memorising the training data, but it is also doing well on the new validation data.

3. Later Phase (50-100 Epochs):

Consistent Low Loss: Towards the end of the training, both the training loss and validation loss remain consistently low. This indicates that the model has learned well and is making very few errors.

No Overfitting: Overfitting happens when the model becomes too good at predicting the training data but fails on new data. In the graph, since the validation loss does not increase and stays close to the training loss, it suggests that our model is well-balanced and not overfitting.

Conclusion from the graph

1. Gradual Improved Learning: The model starts by making a lot of mistakes (high loss), but quickly improves and makes fewer mistakes as it learns (loss decreases).

2. Gradual Balanced Performance: Both the training and validation losses stay low and close to each other, meaning the model is good at both learning from the training data and predicting new data.

3. No Overfitting: The model is not just memorising the training data, it is genuinely learning patterns that help it make good predictions on new, unseen data.

Overall, the graph shows that the model is doing a great job of learning and is ready to make accurate predictions when applied to real-world scenarios.

Solar Energy Panel Installation Potential Map:

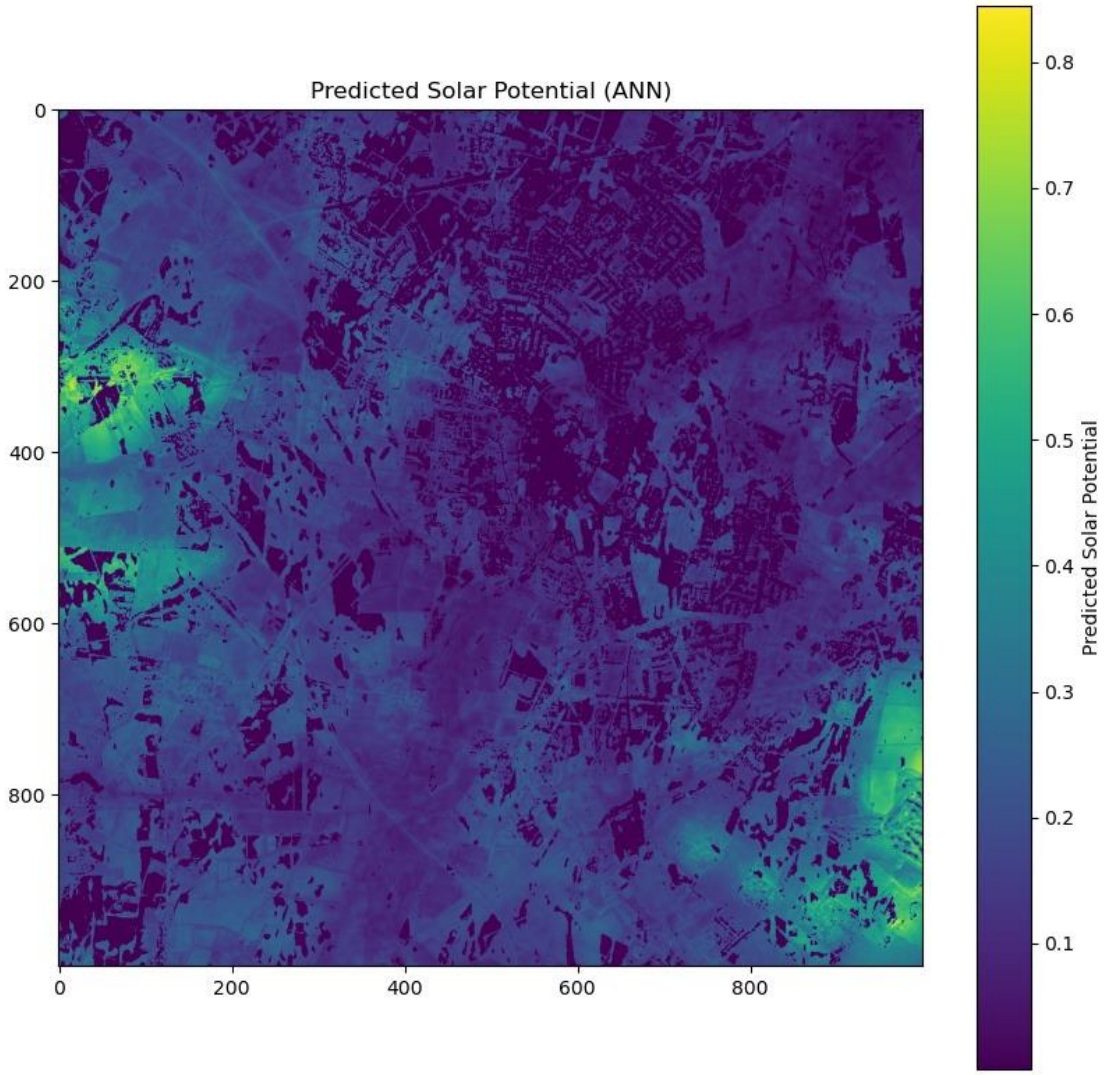


Fig 12: Areas with Solar energy Potential as per ANN

The ANN map shows areas with varying solar potential across the Cambridge region. The brighter (yellow) areas represent higher predicted solar potential, while the darker (purple) areas represent lower potential.

2. Deep Neural Network (DNN)

Like the ANN, the scatter plot for the DNN shows true values against predicted values. The points are also closely aligned with the red line, indicating strong predictive accuracy. This map is expected to provide the most precise prediction of solar potential, making it highly reliable for decision-making.

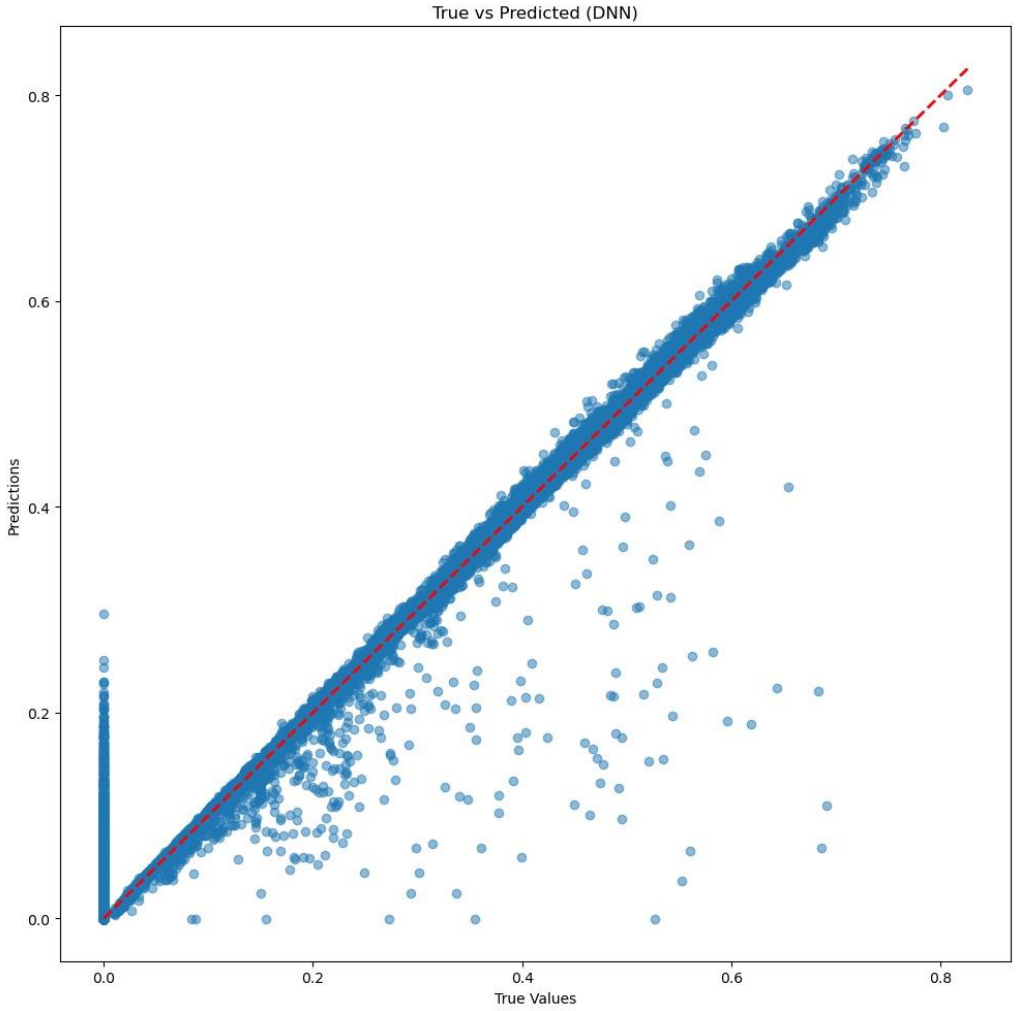


Fig 13: DNN: True V/s Predicted Values

- Mean Squared Error (MSE): 0.00005948844113713921
- R-squared Score (R^2): 0.9959520258020225 which is ~0.996

The DNN model also performs very well, with a slightly higher MSE than the ANN but still very low. The R^2 score of 0.996 indicates that the DNN explains 99.6% of the variance in the data, making it a strong model.

Solar Energy Panel Installation Potential Map:

Like the ANN map, the DNN map also highlights regions with higher and lower solar potential. The distribution of potential is like the ANN map, with some slight variations in intensity.

The DNN model also performed very well, with a slightly higher MSE than ANN but still a high R-squared score. The solar potential map from DNN is nearly as accurate as the ANN map, making it another strong option for predicting solar potential.

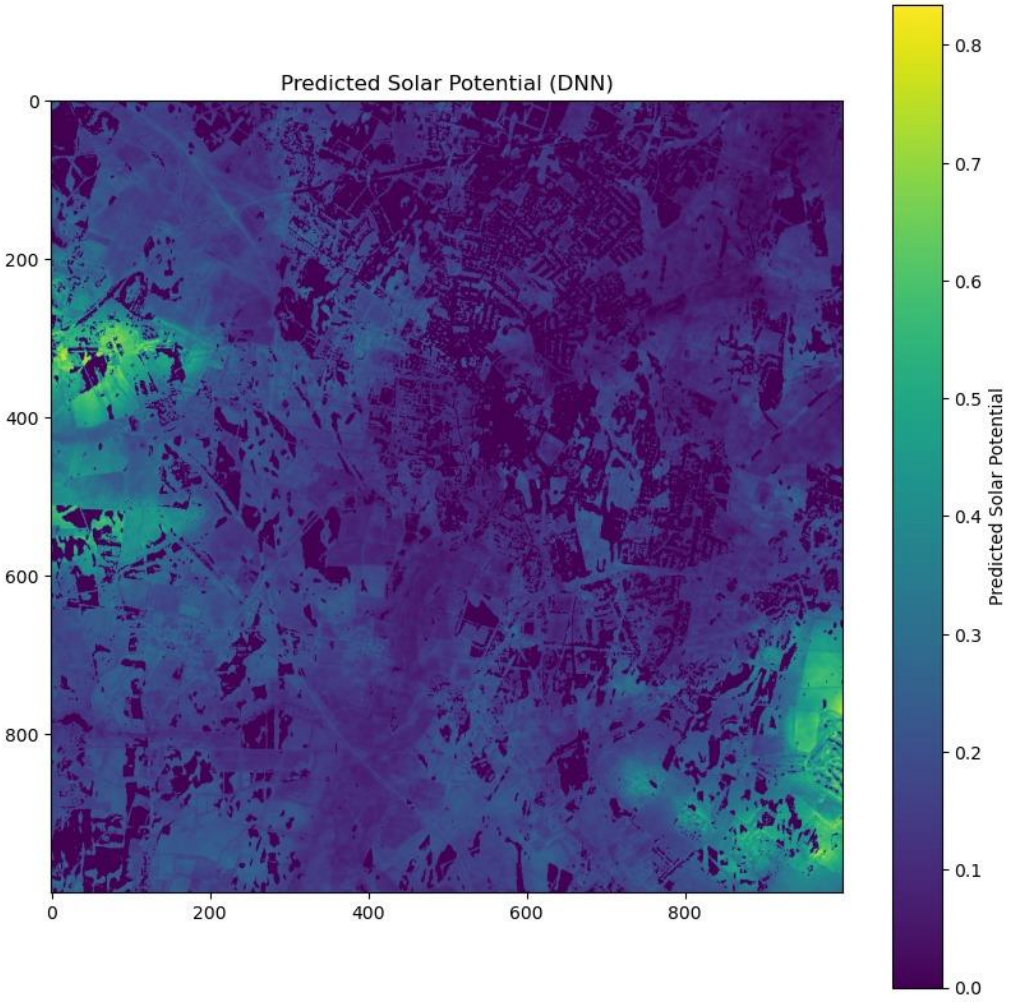


Fig 14: Areas with Solar energy Potential as per DNN

3. Gradient Boosting

The scatter plot for the Gradient Boosting model shows more spread around the red line compared to the neural networks. There is more deviation, especially at the lower and higher ends of the true values.

- Mean Squared Error (MSE): 0.0004428707887100893
- R-squared Score (R^2): 0.9698642376322552 which is ~0.97

The Gradient Boosting model has a higher MSE, indicating more significant prediction errors. The R^2 score of 0.970 is still strong but lower than the neural networks, explaining 97% of the variance. This suggests that though Gradient Boosting is effective, it may not be as precise as the ANN or DNN for this dataset.

Solar Energy Panel Installation Potential Map:

The Gradient Boosting map shows solar potential across the region, with a distribution pattern that is generally consistent with the ANN and DNN maps but with some areas of variation, particularly in the middle range of solar potential.

The Gradient Boosting model had higher MSE and lower R-squared scores compared to the ANN and DNN models. This suggests that while the map provides a good indication of solar potential, it may not be as precise, especially in regions with medium to high solar potential

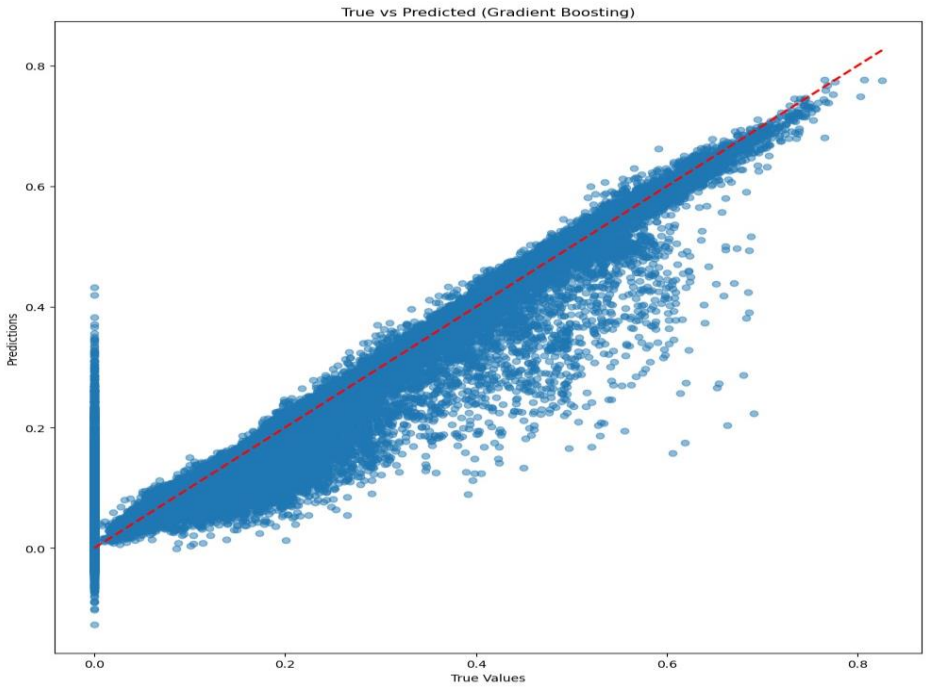


Fig 15: Gradient Boosting: True V/s Predicted Values

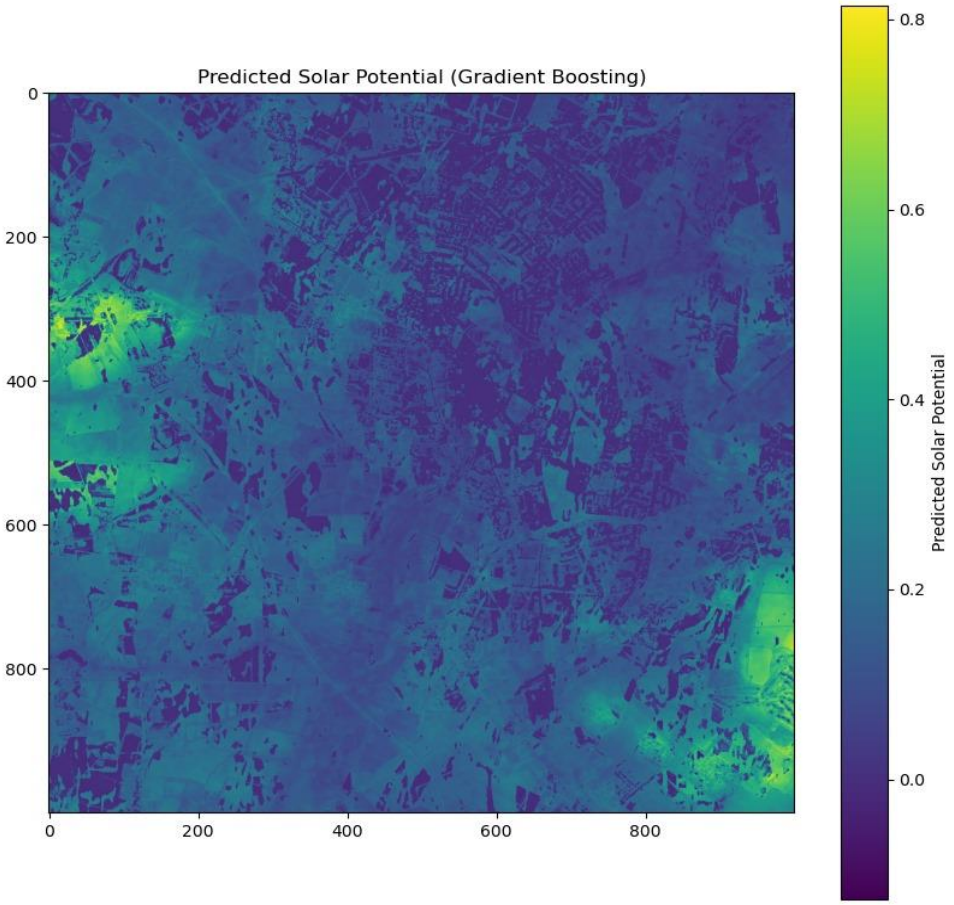


Fig 16: Areas with Solar energy Potential as per Gradient Boosting

4. Random Forest

The scatter plot for the Random Forest model is like the Gradient Boosting, with some spread around the line but less than Gradient Boosting. The predictions generally follow the true values closely.

- Mean Squared Error (MSE): 0.0000817146987296945
- R-squared Score (R^2): 0.9944396090109209 ~0.995

The Random Forest model has a slightly higher MSE than the ANN and DNN, but it is still relatively low. The R^2 score of 0.994 indicates that the model explains 99.4% of the variance in the data, making it a strong contender, though not quite as powerful as the ANN.

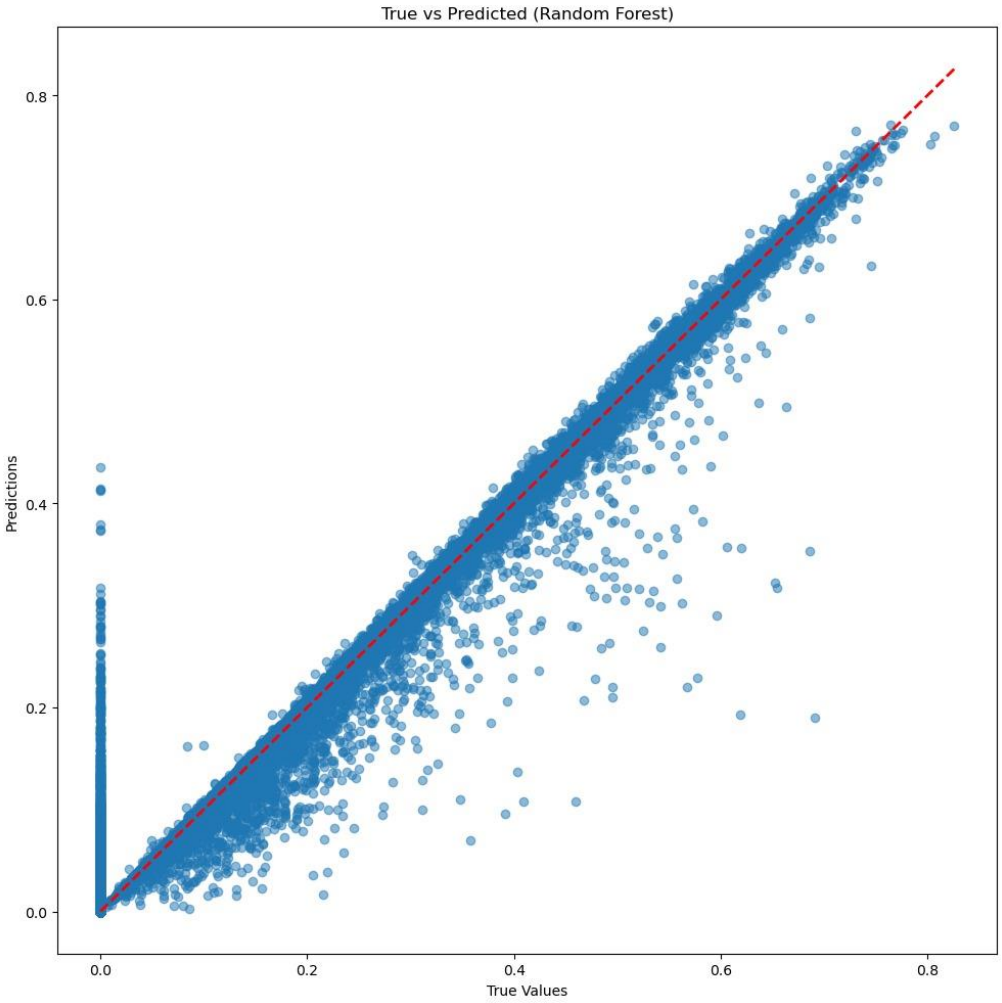


Fig 17: Random Forest: True V/s Predicted Values

Summary of the results

ANN and DNN are the top performers with the lowest errors (MSE) and highest R^2 scores, indicating excellent predictive accuracy and minimal errors. Random Forest also performs well, with a slightly higher MSE than ANN and DNN, but it still explains a significant amount of variance. Gradient Boosting shows more prediction errors with the highest MSE and lowest R^2 among the models but is still a strong model with a solid R^2 of 0.970.

Conclusion and further explorations

Conclusion

The methods explored—especially the ANN and DNN models—have shown excellent predictive capabilities and can be extended to solve a wide range of business problems across various industries. Their commercial applications are vast, from optimising energy usage and financial risk management to improving retail operations and healthcare outcomes. Further exploration of these models, combined with advanced techniques and broader data integration, could unlock even greater value and innovation in these fields.

Further Explorations

Optimising the models:

Tuning the Hyperparameters: While the ANN model has performed best, further improvement could be achieved by fine-tuning hyperparameters such as learning rates, activation functions, and network architecture. This could involve using techniques like grid search or random search to find the optimal parameters.

Ensemble Methods: The outputs of the ANN, DNN and Random Forest, which are the best performing models, may be combined in an ensemble approach. This may yield even better predictive performance by leveraging the strengths of each model.

Additional Data:

Need of Additional Data Sources: Incorporating more diverse datasets, such as historical trends, demographic data, or weather data, could improve the models' robustness and accuracy.

Feature Engineering: Creating new features or transforming existing ones (e.g., polynomial features, interaction terms) could enhance the model’s ability to capture complex patterns in the data.

Historic Model Utilisation:

Applying Pre-trained Models: Using transfer learning, where pre-trained models from similar tasks are fine-tuned on your specific dataset, could reduce training time, and improve performance, particularly for models like DNN.

Recommendations for Stakeholders on Optimal Locations for Solar Panel Installations

Based on the analysis of solar potential using various predictive models, the following recommendations are customised to different stakeholder groups, including policymakers, urban planners, and investors.

1. Policymakers

Maintain and Enhance Incentives for Solar Energy Investment: Enhanced tax incentives and improved locational signals, could attract new investment and alleviate network congestion.

Prioritise High-Potential Regions: We can identify regions in and around the UK using this model which will allow us to focus on promoting solar energy incentives in these regions. The maps generated by the ANN model highlight areas around Cambridge that can yield the highest return on investment in solar infrastructure.

Zoning Regulations: The models can be used as target regions or areas to implement or update zoning regulations that encourage the use of solar panels. Policies could include mandatory solar installations for new buildings or incentives for retrofitting existing structures.

Remove Obstacles to New Solar Energy Capacity: The Winser report provides additional recommendations to accelerate the development of electricity transmission infrastructure. **(Winser, 2023)**

Long-Term Renewable Energy Goals: Using these insights may help us in identifying the best regions to invest for installing solar panels around major towns and cities in the UK. We can use these insights to align with national or regional renewable energy targets. Policymakers can consider these high-potential zones as critical areas to meet renewable energy goals and reduce carbon emissions.

Invest in Storage Solutions, Grid Upgrades, and Essential Grid Services: Without an adequate grid, renewable energy investment will be further constrained, particularly at higher penetration rates. Immediate action, including new incentives for storage, potential market redesign, and the integration of local and national flexibility markets, could attract more private financing and send positive market signals to expedite the energy transition.

2. Urban Planners

Green Infrastructure Development: Consider developing community solar projects in underutilised urban spaces using this model. These projects can provide renewable energy to residents and businesses, particularly in high-density urban areas.

Back-Up/Alternative planning: Integrate solar energy into back-up strategies, ensuring that critical infrastructure in important areas has access to reliable, sustainable power during emergencies or grid failures.

3. Investors

Targeted Investments: Prioritise investments in regions with the highest predicted solar potential, as identified by the ANN and DNN models. These areas offer the best opportunity for maximising returns on solar energy projects.

Diversification of Solar Portfolios: Consider diversifying investments across different high-potential zones to balance risk and take advantage of solar panel installations in lesser known but high potential regions, varying market conditions, grid capacities, and regional incentives.

Long-Term Planning: Invest in large-scale solar farms in rural high-potential areas where land costs are lower, and opportunities for expansive installations exist. These can be paired with investments in battery storage and smart grid technologies to ensure long-term viability.

Partnership Opportunities: Partnerships with local governments and businesses in high-potential regions may be explored to co-develop solar projects. Such collaborations can reduce investment risks and increase project scalability.

Summary

These recommendations are based on predictive models that have identified regions with the highest solar energy potential, particularly in the Cambridge area. By focusing on these optimal locations, stakeholders—including policymakers, urban planners, and investors—can maximise the efficiency, effectiveness, and profitability of solar panel installations. These efforts will not only contribute to the growth of renewable energy but also support broader sustainability goals and energy security.

Future research should focus on ways to accelerate domestic investment in renewable energy and storage solutions. It should also assess the impact of a rapid energy transition in the UK on service and resource demands, including the need for skilled labour and the ethical sourcing of components. Given that resources, technology, and costs are no longer significant barriers, the focus should now shift to addressing the political and economic challenges of an accelerated energy transition.

References

Abdel-Nasser, M., & Mahmoud, K. (2017). Accurate photovoltaic power forecasting models using deep LSTM-RNN. Neural Computing and Applications, 1-14.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Ghafoor, S., & Munir, A. (2021). A review of advances in solar energy systems in the context
Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.

Hoffmann, B., Hamacher, T., & Neumann, D. (2018). The role of flexibility in the light of the integration of renewable energy sources. Applied Energy, 229, 67-89.
https://www.energy.gov/eere/solar/solar-energy-technologies-office

International Energy Agency. (2021). Renewable energy market update. Retrieved from

Mathe, J., Miolane, N., Sebastien, N., & Lequeux, J. (2024). PVNet: A LRCN Architecture for Spatio-Temporal Photovoltaic Power Forecasting from Numerical Weather Prediction.

Murtagh, F., Farid, M., & Carranza, M. (2018). Multivariate and time series analyses for environmental data.

National Renewable Energy Laboratory. (2018). Solar energy potential. Retrieved from
of smart grid. Renewable and Sustainable Energy Reviews, 135, 110-189.

O'Callaghan, B., Hu, E., Israel, J., Smith, C. L., Way, R., & Hepburn, C. (2023, September). Could Britain's energy demand be met entirely by wind and solar?

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

U.S. Department of Energy. (2020). Solar energy technologies office. Retrieved from

World Bank. (2019). Global solar atlas. Retrieved from https://globalsolaratlas.info/

XI, Y., Thinh, N. X., & LI, C. (2019, March 1). Full article: Preliminary comparative assessment of various spectral indices for built-up land derived from landsat-8 oli and sentinel-2a MSI Imageries.
https://www.tandfonline.com/doi/full/10.1080/22797254.2019.1584737

Queen Mary
University of London

qmul.ac.uk