Article

Application of Machine Learning and Deep Neural Visual Features for Predicting Adult Obesity Prevalence in Missouri

Butros M. Dahu [1,2,*], **Carlos I. Martinez-Villar** [3], **Imad Eddine Toubal** [3], **Mariam Alshehri** [3], **Anes Ouadou** [3], **Solaiman Khan** [1], **Lincoln R. Sheets** [1,2] and **Grant J. Scott** [1,3]

1   Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211, USA;
    sknnh@missouri.edu (S.K.); sheetslr@missouri.edu (L.R.S.); scottgs@missouri.edu (G.J.S.)
2   Department of Health Management and Informatics, University of Missouri, Columbia, MO 65211, USA
3   Department of Electrical Engineering and Computer Science, University of Missouri,
    Columbia, MO 65211, USA; cmartinezvillar@mail.missouri.edu (C.I.M.-V.); itoubal@missouri.edu (I.E.T.);
    msalshehri@missouri.edu (M.A.); aomqc@missouri.edu (A.O.)
*   Correspondence: butrosdahu@missouri.edu; Tel.: +1-832-512-4825

Abstract: This research study investigates and predicts the obesity prevalence in Missouri, utilizing deep neural visual features extracted from medium-resolution satellite imagery (Sentinel-2). By applying a deep convolutional neural network (DCNN), the study aims to predict the obesity rate of census tracts based on visual features in the satellite imagery that covers each tract. The study utilizes Sentinel-2 satellite images, processed using the ResNet-50 DCNN, to extract deep neural visual features (DNVF). Obesity prevalence data, sourced from the CDC's 2022 estimates, is analyzed at the census tract level. The datasets were integrated to apply a machine learning model to predict the obesity rates in 1052 different census tracts in Missouri. The analysis reveals significant associations between DNVF and obesity prevalence. The predictive models show moderate success in estimating and predicting obesity rates in various census tracts within Missouri. The study emphasizes the potential of using satellite imagery and advanced machine learning in public health research. It points to environmental factors as significant determinants of obesity, suggesting the need for targeted health interventions. Employing DNVF to explore and predict obesity rates offers valuable insights for public health strategies and calls for expanded research in diverse geographical contexts.

Keywords: DCNN; geospatial; machine learning; obesity rate; satellite imagery

1. Introduction

This research study addresses the escalating obesity epidemic in the United States and globally, emphasizing its complex relationship among various factors. It highlights the significant advancements in understanding this issue through deep learning and machine learning techniques, particularly the analysis of large satellite imagery datasets (commonly called Earth Observation). It also focuses on the discrepancies in previous studies and introduces a novel methodology using deep convolutional neural networks (DCNN) to extract features from satellite images, combined with an advanced machine learning regression model to predict obesity prevalence from satellite images. This approach aims to provide a comprehensive and scalable analysis of the correlation between deep neural visual features (DNVFs) and obesity rates utilizing publicly accessible data.

Obesity has escalated to an epidemic level across the United States [1]. From 2001 to 2018, there has been a steady rise in obesity rates among adults over 20 years old [1]. The Global Burden of Disease study reveals that over 603 million adults were obese globally in 2015 [2,3]. In the United States, over a third of adults are obese, with 46 states reporting an adult obesity rate of at least 25% [2,4]. Obesity is a complex issue influenced by various factors such as the higher risk of various non-communicable diseases [5,6], including cancer,

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

2 of 23

mobility issues, heart disease, mental health problems, osteoarthritis, sleep apnea, stroke, and Type 2 diabetes [1,2,7].

Additionally, obesity causes an estimated 300,000 deaths annually, just behind smoking-related deaths [1,8]. Unhealthy eating habits and inactive lifestyles are often linked to aspects of both the social and built environments, encompassing natural and modified physical surroundings [7]. It impacts health by dictating the availability of resources like housing, spaces for activities and recreation, and community design features [2,9–12].

The growing prevalence of obesity and related health issues in urban settings has led researchers to delve into the complex relationship between public health and individuals' surroundings, which makes using raw overhead satellite images to predict a health metric an interesting approach [13–16]. This exploration has been significantly advanced by the application of deep learning and machine learning techniques, which have enabled the analysis of large datasets from diverse sources such as satellite and street view imagery [17–21].

Previous research indicates a connection between various aspects of the built environment and their impact on obesity and physical activity throughout different phases of life [7,22–24]. Previous studies also highlight a link between obesity and several environmental variables such as the walkability of an area, its land use patterns, the extent of urban sprawl, the type of residential area, availability of resources like recreational facilities and food establishments, the degree of socio-economic deprivation, and the level of perceived safety in an area [25–28]. Moreover, being close to and having access to natural spaces and sidewalks is associated with higher and more consistent physical activity levels, particularly in urban environments [29–31].

Although there is a recognized relationship between obesity and the built environment, there have been noted discrepancies in the findings of different studies and across various geographical areas regarding the impact of specific built environment features on obesity rates [25,32–34]. These inconsistencies might stem from differences in the methods and tools used for measurement in these studies, which complicates the evaluation and comparison of the results [35–37]. Additionally, measuring these environmental features often involves significant costs and time, and is prone to human error and bias [38–40]. There is a need for methodologies that offer uniform measurement standards to facilitate comparisons across different studies [36,41,42]. Accurately determining the influence of the built environment on obesity is important for the development and execution of effective community-based prevention and intervention strategies [38,43–45].

Herein, we present a novel methodology to thoroughly investigate the link between the prevalence of adult obesity using solely overhead satellite images. This method utilizes a deep learning technique, specifically a DCNN is used to analyze the physical characteristics of neighborhoods from medium-resolution (10 m) satellite images. This approach builds on the work of Maharana et al. [2] and Nguyen et al. [46], who employed DCNNs to categorize images from Google Street View, examining the relationship between obesity and specific elements like crosswalks, building types, and the presence of greenery or landscaping. However, their research was constrained to these three preselected features and did not fully leverage the potential of DCNNs to identify features correlated with obesity rates autonomously. Our method, in contrast, offers a more comprehensive analysis of overhead image features, pinpointing detailed correlations with obesity rates at the census tract level in 1052 census tracts across the State of Missouri in the United States. Additionally, our approach is scalable, utilizes publicly accessible data and computational resources, facilitates comparison between studies, and can be applied to different geographic sites and regions.

The objective of this study is to employ deep convolutional neural networks (DCNNs) to analyze medium-resolution satellite images and identify physical characteristics of neighborhoods that correlate with obesity rates. By focusing on the 1052 census tracts in Missouri, USA, the study offers a more comprehensive and scalable approach to assessing the impact of deep neural visual features on obesity prevalence, enhancing the precision of public health interventions and policy-making. The importance of predicting obesity using

Int. J. Environ. Res. Public Health **2024**, 21, 1534

3 of 23

deep neural network features is underlined by our method's ability to enhance the precision of public health interventions and improve the granularity of epidemiological studies.

This paper is structured as follows. In Section 2, we provide an overview of the related literature. Section 3 outlines the methods employed in this study, followed by a discussion of the results in Section 4. In Section 5, we discuss the findings' implications and acknowledge our work's limitations. Finally, Section 6 presents the overall conclusions and key takeaways.

2. Related Work

Songhyeon SH et al. [47] and Lam TM et al. [48] showcase the diverse methodologies employed in the field. They used a combination of multinomial logistic regression and Deep Neural Network algorithms to analyze community health survey data from 2018 to 2020. The study underscores integrating statistical methods with advanced machine learning techniques in public health research. Lam TM et al. [48] conducted an umbrella review, synthesizing evidence from 32 systematic reviews to evaluate the connection between the built environment and obesity. This meta-analysis focused on the methodological aspects of the reviewed studies, particularly assessing biases and thematic areas, thus providing a comprehensive overview of research trends and methodological approaches in this domain.

The studies by Alkhalaf M et al. [49] and An R et al. [50] delve into the advanced applications of machine learning and artificial intelligence in obesity research. Alkhalaf M et al. conducted a comprehensive analysis of ML applications in adult obesity studies, focusing on the evaluation of various algorithms, including regression models, neural networks, and deep learning techniques [49]. R. An et al. assessed the technical aspects of AI models, particularly ML and DL, in obesity research, analyzing 46 studies that utilized a range of AI methodologies [50]. Zhou et al. [19,51] explored the use of machine learning models in the context of obesity research, with a focus on biomarker detection and intervention strategies. Their review categorized ML models into supervised and unsupervised learning types and detailed 25 open-source ML algorithms, platforms, and databases relevant to various aspects of obesity research.

DCNNs have seen emerging use in population health studies, across various image and sensing modalities and tasks [52–54]. Maharana et al. study utilized the VGG-CNN-F model to analyze approximately 150,000 high-resolution satellite images [2]. Newton et al. [55] employed the Xception DCNN architecture, allowing for a more efficient process than traditional DCNNs. Yue X et al. [56] and Phan L et al. [57] both emphasized the use of CNNs for analyzing neighborhood characteristics, but with different approaches and scales. They employed VGG19 [58] and ResNet18 [59] architectures to analyze a vast dataset of 164 million Google Street View images. The study highlighted the effectiveness of using different CNN architectures to manage and interpret large-scale image data. Phan L et al. [57], on the other hand, focused on a dataset of 31,247,167 Google Street View images, utilizing the VGG-16 model. Their objective was to evaluate built environment indicators at the state level in the United States and explore their association with public health outcomes.

3. Materials and Methods

The study analysis consisted of three steps. First, we processed Sentinel-2 satellite images to extract features of any environment using ResNet-50 [59]. Second, we merged the census tract polygons from Tiger Line with the polygons from the CDC data to match each census tract with its respective obesity rate [60]. Third, we used a machine learning model, generalized linear model (GLM), random forest, and 10-fold-cross validation to build a parsimonious model to predict the obesity rate for each census tract and to assess the association between the built environment and obesity prevalence. The study was exempt from institutional review board approval because this research used existing data and records collected by external parties in such a manner that individuals cannot be identified.

Figure 1 summarizes the methodology of our research study, which leverages remote sensing and machine learning to estimate regional obesity rates. Starting with medium-

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

4 of 23

resolution images from Sentinel-2, the study utilizes the ResNet-50 to extract pertinent DNVF from the satellite data. These features, encoded as a 2048-dimensional vector or concepts embedding, contain critical visual information that reflects the environment's characteristics, potentially linked to obesity prevalence.

The extracted features then undergo regression analysis within a machine learning framework to predict the estimated obesity rates for the study area. This statistical approach models the relationship between the DNVF and the obesity rates, providing an innovative angle to public health research.
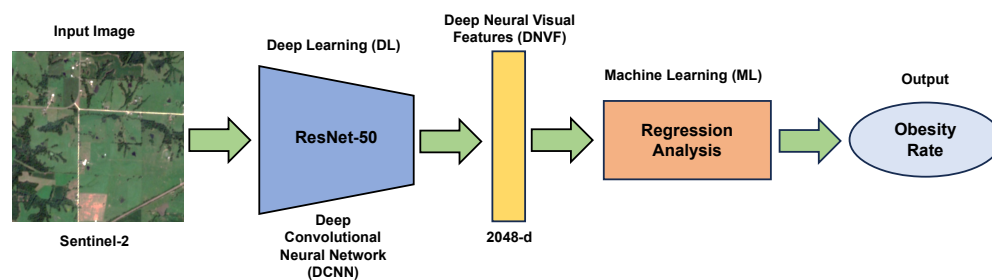

Figure 1. Flowchart illustrating the estimation of obesity rates from satellite imagery using a combination of deep learning with ResNet-50 architecture and machine learning regression analysis.

3.1. Obesity Prevalence Data

We utilized 2022 estimates of annual crude obesity prevalence at the census tract level, derived from the 500 Cities project (PLACES: Local Data for Better Health, accessed on 17 March 2024). These estimates are based on data from the Behavioral Risk Factor Surveillance System, which surveys individuals aged 18 and older. Obesity is identified using a body mass index (BMI) threshold of 30, calculated as the individual's weight in kilograms divided by their height in meters squared [61]. Our study focused on the Mid-Missouri region in the United States. The 1052 census tracts (Missouri State) covered in this study have an aggregate area of 69,707 square miles (180,540 square km). They have a total population of 6.2 million (based on 2020 census).

The CDC data lists 1387 census tracts, 4506 block groups, and 343,565 census blocks. Given that the number of Tiger Line census tract shapefiles in the state was 1654, polygon (census tract) IDs in the Tiger Line data had to be aligned to census tract IDs in the CDC data. To fix the mismatching issue, census tracts with subdivisions (tract names with two trailing digits different from zero) in both datasets were joined into larger polygons.

First, we joined the polygons in the Tiger Line dataset by removing all subdivisions. Of the initial 1654 polygons, 914 had names following a naming convention of the type "XXXX.YY", where the "YY" corresponded to the subdivision within a particular tract. The remaining 740 had no subdivisions and names containing two trailing zeros (e.g., "XXXX.00"). Setting the two trailing digits to zero in these 914 names resulted in repeated names with 323 unique names. If a name was repeated, all repeated elements were joined to become a single polygon, which resulted in a set of 1063 polygons.

Similarly, the IDs of the census tracts in the CDC data consisted of a string of digits, with the last two digits corresponding to a set of subdivisions different from that of the Tiger Line data. Of the 1387 tracts, 881 had IDs with two trailing zeros (no subdivisions), while the remaining 506 had subdivisions. We set the initial 506 string IDs to a single subdivision with repeated entries, which resulted in 178 unique IDs. Since each of these new unique IDs had multiple obesity rates, an average obesity rate was calculated as a proxy for the obesity rate of the newly joined area. The average obesity rate for these joint areas, weighted by census tract population, was calculated as

$$\bar{w} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i},$$ (1)

Int. J. Environ. Res. Public Health **2024**, 21, 1534

5 of 23

where $\bar{w}$ is the new obesity rate for the joint area, n is the number of subdivisions (repeated entries) within the original census tract, w_i is the population in a subdivision i, and x_i is the obesity rate in i. The joined 1059 CDC entries were then matched to the joined 1063 Tiger Line tracts, for a final overlapping set of 1055 census tract polygons with their corresponding obesity rates, which we used as the inputs to our models.

3.2. Acquiring Satellite Imagery

We determined our image inputs by selecting the Sentinel-2 products intersecting our previously defined set of census tracts. These satellite products were downloaded from ESA's Copernicus Dataspace Ecosystem. Since ESA's OpenSearch API uses HTTP requests to search for products, we defined a shortened geometry string that could fit in our search query. This was done by joining our previously defined census tract polygons into a state boundary that was further simplified into a closed polygon of 54 vertices using the implementation of the Douglas-Pecker algorithm included in GeoPandas.

Our search resulted in 187 intersecting Sentinel-2 products between 1 July 2022, and 31 August 2022. Overlapping images were removed in two steps. First, products with completely overlapping geometries (corresponding to the same UTM zone tile) were filtered by discarding all but the product with the largest area and the lowest cloud percentage. Second, seven partially overlapping products (which also happened to have little state coverage) were discarded after visual inspection. This resulted in a set of 33 Sentinel-2 images that were used to define our inputs to the neural network. These products were downloaded from ESA's Dataspace Ecosystem to Nautilus. All 33 Sentinel-2 image sizes were 10,980 by 10,980 pixels. The images were then normalized to values between 0 and 1 and cropped into chips of 224 by 224 pixels. This created a total of 82,500 three-band (RGB) image chips.

3.3. Image Processing

DCNNs (e.g., ResNet-50) have made significant strides in various computer vision tasks such as object detection and image segmentation, especially when dealing with extensive data sets [3,62]. These advances are not only crucial in general technology fields but also have profound implications in healthcare, such as in the identification of skin cancer, and social issues like poverty estimation [4,63]. Due to the absence of a substantial labeled satellite dataset–with sufficient data to train a neural network–for categorizing regions with high and low obesity rates, we used the DNVF from a pre-trained neural network to obtain the built area features of the 82,500 satellite image chips. We intersected our 82,500 chips with our Missouri census tracts which results in 63,592 chips that were usable. Only chips that intersected census tracts and were fully within the joint polygon of the state of Missouri were and everything else was discarded.

We used the ResNet-50 network, which is composed of 50 layers (48 convolution layers along with one max pool and one average pool layer) and is trained on approximately 1.2 million images from the ImageNet database (a data set of >14 million images used for large-scale visual recognition challenges) [64] for recognizing objects belonging to 1000 categories [64]. For each chip we passed through the network, we extracted the 2048 features from the last hidden layer of the network before the output layer [65]. Before passing through, each image is standardized using the ImageNet per-band mean and standard deviation. This resulted in a final dataset–used for our machine learning analysis of 1051 rows and 2048 columns, with each row corresponding to the average feature vector of a census tract. Since each census tract in our dataset could have more than one image chip intersecting it, we calculated a corresponding weighted mean feature vector for a tract with the features of the intersecting chips [66]. For a census tract t, its mean feature vector F_t was calculated as:

$$F_t = \frac{\sum_{\forall c \in \mathcal{C}} w_c F_c}{\sum_{\forall c \in \mathcal{C}} w_c}, \qquad (2)$$

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

6 of 23

where $\mathcal{C}$ is the set of image chips intersecting the census tract, F_c is a ResNet-50 feature vector obtained from chip c, and w_c is the (scalar) number of pixels in c intersecting the census tract.

To include the obesity rate of a given chip in our analysis, the inverse case was considered. When a chip intersected with more than one census tract, we calculated the obesity rate in such a chip as the average obesity rate of all the intersecting census tracts [66]. So, for a chip c the weighted obesity rate is similarly calculated as:

$$o_c = \frac{\sum_{\forall t \in \mathcal{T}} w_t o_t}{\sum_{\forall t \in \mathcal{T}} w_t}, \tag{3}$$

where $\mathcal{T}$ is the non-empty set of tracts intersecting a chip c, o_t is the obesity rate in a census tract, and w_t is the number of pixels in c corresponding to the t census tract. We do not link these features to specific elements in the built environment or the obesity prevalence. Rather, these DNVFs collectively represent an indicator to help predict the obesity prevalence for each census tract.

The image chips were cropped for a size of 224 pixels by 224 pixels or 2240 m (2.24 km) by 2240 m (2.24 km). The GPU-accelerated processing speed was approximately 77.51 tiles per second, for a total of 17.73 min. This translates to a rate of 0.0129 s per tile.

Figure 2 shows our geospatial data analysis using 33 Sentinel-2 satellite imagery within the state of Missouri for the year 2022. Figure 2A on the left of the figure illustrates a map of Missouri, demarcated by latitude and longitude, overlaid with 33 distinct green rectangles signifying the spatial coverage of Sentinel-2 images. These images encapsulate notable geographic demarcations including county boundaries and principal urban areas. The central portion of the figure describes the process undertaken to standardize the raw satellite data, which entails normalizing the images to ensure homogeneity in terms of color and scale across all datasets. After this normalization, the images are methodically cropped and the size of each image chip is equal to 224 pixels by 224 pixels or 2.24 km by 2.24 km. These chips are superimposed on blue outlines that delineate 1052 census tracts of our study, suggesting a systematic intersection of medium-resolution satellite data with granular demographic units.
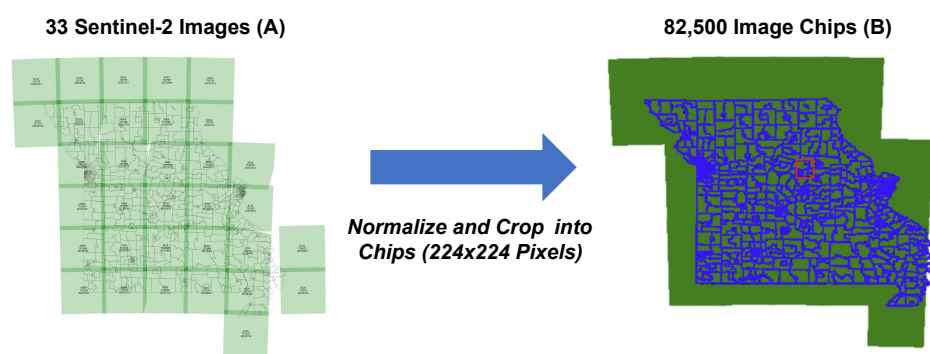

Figure 2. Data processing workflow for Sentinel-2 satellite imagery within Missouri in 2022. (**A**) Displays the geographic coverage of 33 Sentinel-2 images across Missouri, with county boundaries. The central diagram outlines the normalization process and the cropping of images into 224 × 224 pixel chips. (**B**) Illustrates the distribution of 82,500 resultant image chips. The red box represents Mid-Missouri area and Boone County.

Figure 2B on the right provides a detailed view of the Sentinel-2 image coverage as it intersects with the state-designated census tracts. It presents the state of Missouri as a green plane, superimposed with an array of green squares, each representing the 82,500 image chips generated post-normalization. The distribution of these image chips across the state varies, with a higher density in certain regions, potentially reflecting areas of research study. This meticulous preprocessing routine is essential to prepare the data for the subsequent

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

7 of 23

analytical procedures, ensuring precision and reliability in the results of the spatial analysis conducted within this research study.

3.4. Regression Analysis of Predictive Capability

We used the 10-fold cross-validation which is a widely used method for evaluating the performance of a model. It involves dividing the dataset into ten equal parts, or "folds". The model is trained on nine folds and tested on the remaining one. This process is repeated ten times, with each fold serving as the test set once. The results from all ten iterations are averaged to provide an overall assessment of the model's performance. This approach helps to minimize bias and variance in the evaluation, ensuring a more reliable and generalizable estimate of the model's effectiveness [67–69].

In the evaluation of the 10-fold cross-validation fit we utilized three key metrics: Mean Squared Error (MSE), R^2, and Adjusted R^2. MSE is a measure of the average squared difference between the actual observed outcomes and the outcomes predicted by the model, providing a clear quantification of the model's prediction error [70]. Lower MSE values indicate better model accuracy. R^2, also known as the coefficient of determination, assesses the proportion of the variance in the dependent variable that is predictable from the independent variables [71]. It offers an insight into the goodness of fit of the model, with values closer to 1.0 suggesting a better fit [71]. Adjusted R^2 adjusts for the number of predictors in the model, providing a more accurate measure of the goodness of fit, especially important in models with a high number of predictors [72]. These metrics collectively offer a comprehensive evaluation of 10-folds cross-validation model performance.

4. Results

Figure 3 shows a choropleth map of obesity prevalence among the various census tracts within the state of Missouri for the year 2022. In the map, lighter shades of red are lower obesity rates, and darker hues correspond to higher obesity rates. The scale itself delineates a range starting at 25%, represented by a light red, and progresses to 50%, indicated by a dark red color. Notably, the map reveals a significant variation in obesity rates across different regions, with some census tracts exhibiting markedly higher rates and others reflect lower obesity rates.
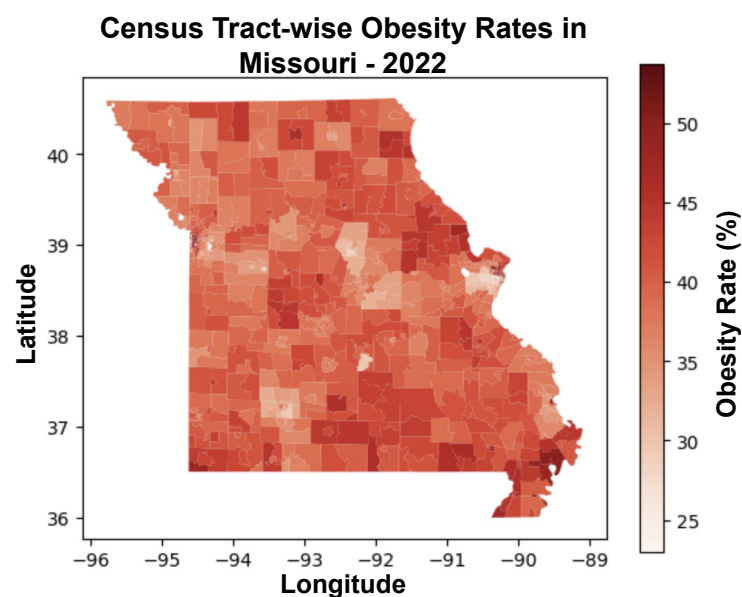

Figure 3. Choropleth map displaying the distribution of obesity rates percentage for individuals across Missouri census tracts in 2022. The variations in the color intensity reflect the range of obesity prevalence, with darker red indicating higher obesity rates. The color scale to the right quantifies the obesity rates corresponding to each color shade.

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

8 of 23

Table 1 presents a detailed summary of key metrics across various regions such as population, area (in square Km), obesity crude prevalence (%), and number of chips. The data in the table includes the minimum, median, and maximum values for each metric, providing a clear understanding of the range and central tendencies within the dataset. For population, the minimum value recorded is 102, the median is 4058, and the maximum reaches 75,569, reflecting the diversity in population sizes across different areas. The area of the regions varies significantly, with the smallest being just 0.49 square Km, the median at 20.19 square Km, and the largest extending to 1787.47 square Km. In terms of obesity rates, the lowest rate observed is 23%, the median stands at 39.20%, and the highest rate is 53.7%, indicating varied health metrics across the regions. Finally, the number of chips ranges from a minimum of 1 to a maximum of 442, with a median of 14, highlighting different levels of chip distribution or consumption.

Table 1. A summary of key regional metrics, covering total population, area (square Km), obesity crude prevalence, and number of chips intersected with each tract.

	Population	Area (Sqr. Km)	Obesity Prevalence (%)	Number of Chips
Min	102	0.49	23.00	1
Median	4058	20.19	39.20	14
Max	75,569	1787.47	53.70	442

In Figure 4, highlights the spatial relation between our 63,592 image chips and the 1052 census tracts within the state of Missouri. Figure 4A provides an overview of the state of Missouri along with two examples of the chip and census tract spatial relationship complexity, marked by a blue boundary indicative of regional limits such as census tracts boundary. The green grid lines (squares/tiles) represent the image chips, which serve as discrete units of spatial data collection or observation points. Figure 4B Zoomed-in view around the Boone county area, narrows the focus to a smaller region, providing a more detailed look at the grid alignment of the image chips with respect to Boone County's geography. A red square highlights a specific area within Boone County, presumably to denote an area of special interest or higher detail study. This zoomed-in perspective allows for an appreciation of the granularity of the image chip distribution within the context of finer geographical boundaries.
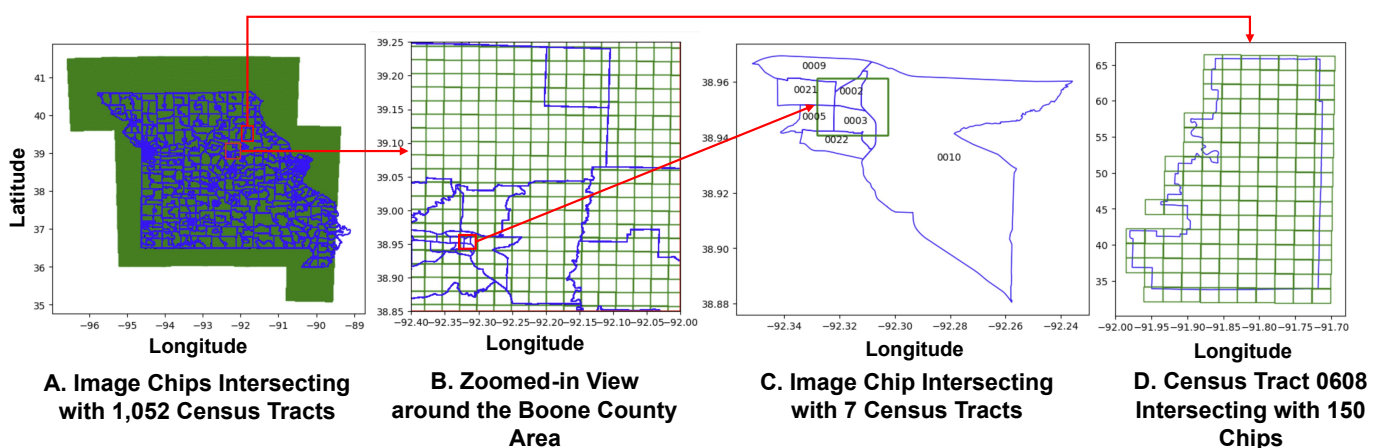

A. Image Chips Intersecting with 1,052 Census Tracts

B. Zoomed-in View around the Boone County Area

C. Image Chip Intersecting with 7 Census Tracts

D. Census Tract 0608 Intersecting with 150 Chips

Figure 4. Multiscale analysis of satellite image chips and census tracts in Missouri. (**A**) exhibits a statewide view with image chips overlaying 1052 census tracts, indicating extensive data coverage. (**B**) zooms into the Boone County area, detailing the alignment of image chips to local geography. (**C**) details individual image chips boundaries, illustrating their overlap with seven distinct census tracts (numbered for reference). (**D**) further narrows down to Census Tract 0608, demonstrating the intersection with 150 specific image chips for granular analysis. The figure highlights the granularity and density of data distribution within the geographic study area.

Int. J. Environ. Res. Public Health **2024**, 21, 1534

9 of 23

Figure 4C shows individual image chip that intersects seven distinct census tracts. This visualization captures the specific overlap between the selected image chip and seven distinct census tracts, allowing for a focused examination of these interactions. This image chip intersects with one of the highest number of census tracts (0002, 0003, 0005, 0009, 0010, 0021 and 0022) and it is located in Boone County. Figure 4D is a single census tract, specifically "Census Tract 0608", with 150 intersecting image chips. This census tract is among the areas with one of the highest number of image chips. This level of granularity reveals a concentrated cluster of data points, potentially signifying a region of particular interest or higher measurement intensity.

Overall, the progression from a broad overview to a detailed view in the four figures systematically demonstrates the methodological approach of correlating medium-resolution satellite imagery with detailed census tract data. Such a multi-scale visual representation aids in understanding the spatial analysis techniques used in our research study, highlighting how local demographic characteristics may be inferred or validated through the meticulous overlay of image chips onto census tract maps. This meticulous approach underscores our research's emphasis on precision and spatial specificity in its analysis.

Figure 5 presents a series of scatter plots, each corresponding to one of ten folds from a 10-fold cross-validation procedure, used to assess the performance of a predictive model for obesity rates. Cross-validation is a machine learning method used to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.
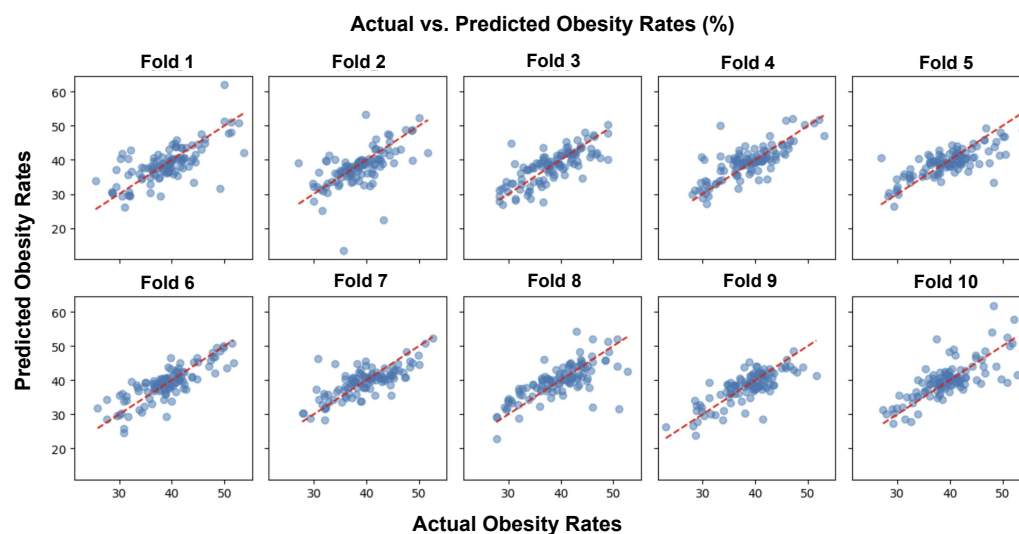

Figure 5. Scatter plots show the relationship between actual and predicted obesity rates using a GLM machine learning model across 10 distinct cross-validation folds. The red dashed line represents perfect prediction accuracy.

In every individual plot, the horizontal axis represents the actual obesity rates (%) and the vertical axis represents the obesity rates as predicted by the model (%). Each plot is labeled as "Fold 1" through "Fold 10", indicating the sequential partitioning of the dataset used for validation. In all plots, the red dashed line represents the theoretical perfect prediction.

A visual inspection across all ten folds shows that the model's predictions are reasonably well aligned with the actual data, as indicated by the cluster of dots around the linear trend line. However, there is variability among the folds, with some showing tighter clustering around the line (indicating more accurate predictions) and others showing more spread (indicating less accurate predictions). This is expected in cross-validation due to the random partitioning of data into different folds, and it allows for the assessment of the model's robustness and generalizability.

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

10 of 23

The uniformity in the distribution of points across the ten folds indicates that the model is consistently making predictions with a similar degree of accuracy. There is no single fold that appears to be an outlier in terms of prediction quality, which suggests that the model is stable and performs equally well across different subsets of the data.

4.1. Generalized Linear Model Regressor (GLM)

Figure 6 shows the actual obesity rates against those predicted by the GLM. The horizontal axis denotes the actual obesity rates derived from empirical observations, while the vertical axis corresponds to the predicted obesity rates yielded by the machine learning model. Each point on the plot represents a single census tract, with its position reflecting the actual rate on the x-axis against the model's prediction on the y-axis. The red dashed line runs through the data points, indicating the linear regression model's line of best fit. This line encapsulates the general direction that the model attributes to the relationship between the actual and predicted values of obesity rates. In an ideal scenario where predictions perfectly match the actual rates, all data points would align precisely along this line.

Figure 6. Scatter plot displaying the relationship between actual and predicted obesity rates using Generalized Linear Regression (GLM), illustrating a moderate degree of correlation with an R^2 value of 0.44 and an adjusted R^2 of 0.43. The close fit is further evidenced by a moderate Mean Squared Error (MSE) of 18.64. These metrics are provided to assess the accuracy of the model predictions.

In the process of evaluating the predictive performance of GLM for obesity rates, the dataset was partitioned into two subsets; 80% of the data was used for training the model, ensuring it could learn the underlying patterns and relationships effectively. The remaining 20% (selected at random) constituted the testing set, which provided a separate and unbiased evaluation of the model's predictive accuracy. This approach is critical for validating the model's ability to generalize to new, unseen data and for minimizing the risk of overfitting. Figure 6 is annotated with key statistical metrics that quantify the model's predictive performance. The model is able to moderately predict the obesity rate from the satellite imagery with a mean squared error (MSE) of 18.64, an R^2 equal to 0.44, and an adjusted R^2 value equal to 0.43. The plot and the metrics together provide a concise summary of the model's performance and highlight areas where further model refinement or additional data collection might be necessary to improve the accuracy of the obesity rate predictions.

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

11 of 23

4.2. Random Forest (RF)

Figure 7 presents a scatter plot visualizing the comparison between actual and predicted obesity rates using a Random Forest regression model. The horizontal axis represents the actual obesity rates expressed as a percentage, while the vertical axis represents the predicted obesity rates, also in percentage terms. Each blue dot on the plot represents a data point where the actual and predicted obesity rates for a given observation are plotted against each other.
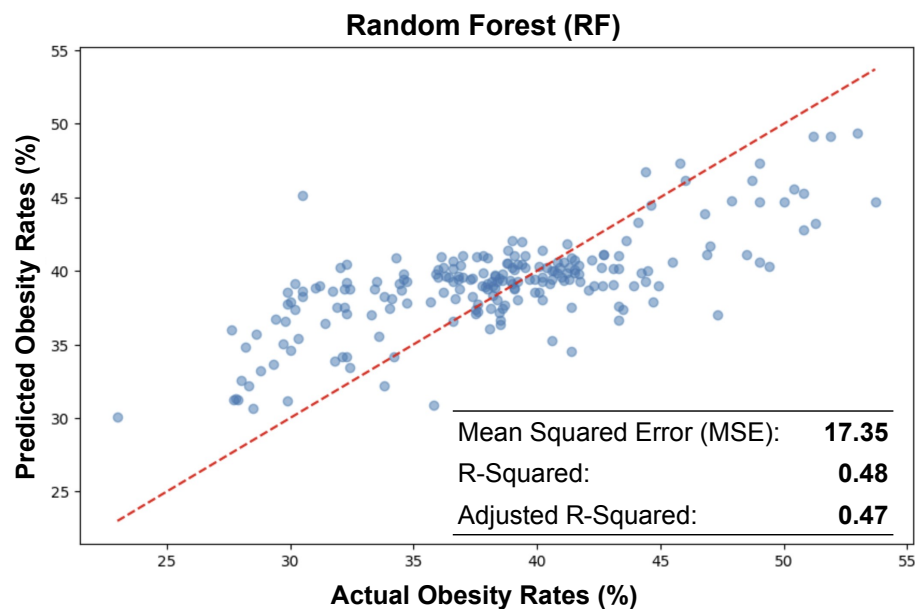

Figure 7. Scatter plot of the relationship between actual and predicted obesity rates using random forest, illustrating a moderate correlation with an R^2 value of 0.48 and an adjusted R^2 of 0.47. The close fit is further evidenced by a moderate Mean Squared Error (MSE) of 17.35.

The red dashed line represents the line of best fit, illustrating the relationship between the actual and predicted values. Ideally, if the predictions were perfect, all dots would fall on this line, which would indicate a 1:1 correspondence between actual and predicted rates. However, the spread of the dots around this line suggests variability in the model's accuracy, with some predictions being more accurate than others.

In our study, we utilized a Random Forest algorithm to forecast obesity rates across various census tracts. To ensure a robust assessment of the model's predictive power, we divided our dataset into two distinct portions; 80% was allocated for training the model, allowing it to learn from a substantial majority of the data, while the remaining 20% (selected at random) was set aside for testing purposes. This testing set served as a crucial benchmark for evaluating the model's efficacy in accurately predicting obesity rates on new, unseen data, thereby safeguarding against overfitting.

The inset box at the bottom of the plot provides key statistical metrics evaluating the performance of the model. The Mean Squared Error (MSE) is reported as 17.35, which gives the average of the squares of the errors that is, the average squared difference between the estimated values and the actual value. The R^2 value is 0.48, suggesting that approximately 48% of the variability in the actual obesity rates can be explained by the model. The Adjusted R^2 is 0.47, slightly lower than the R^2, which takes into account the number of predictors in the model and provides a more adjusted estimation of the goodness of fit.

The positioning of the dots in relation to the line of best fit and the reported R^2 values suggest that while the model has a moderate predictive power, there is a significant portion of the variance in obesity rates that is not captured by the model. This could be due to the complexity of factors affecting obesity rates that are not included in the model or due to the inherent limitations of the Random Forest algorithm when applied to this particular dataset. As we mentioned previously for GLM, the RF plot and metrics together also provide a

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

12 of 23

concise summary of the model's performance and highlight areas where further model refinement or additional data collection might be necessary to improve the accuracy of the obesity rate predictions.

4.3. Features Importance

Table 2 provides a breakdown of the top ten features in terms of their importance percentages, derived from the 10-fold-cross validation model aiming to predict the obesity rates using Deep Neural Visual Features (DNVFs). Feature 1112th is the most significant, holding an importance of 15.01%, indicating its strong influence on the model's predictions regarding obesity rates. The subsequent features, although less important than Feature 1112th, still contribute notably to the model's prediction and include Features 0095th and 1314th, with importance percentages of 3.23% and 2.47%, respectively.

Table 2. Lists the top ten visual features ranked by their importance in our obesity rate prediction model. The percentages indicate the relative importance of each feature in influencing the model's predictions, with higher values signifying greater influence. The values of the correlation coefficient, highlighting the magnitude of each feature's impact.

Feature #	Importance (%)	Correlation Coefficient
1112	15.01	−0.6061
0095	3.23	−0.2196
1314	2.47	−0.2079
0767	2.07	0.1542
0239	2.05	−0.1526
1253	1.05	0.1140
0895	1.04	0.0932
1126	0.87	−0.0536
0338	0.84	−0.0434
0668	0.75	−0.0061

The coefficients in the third column indicate the strength and direction of the relationship between the identified features and obesity prevalence. Positive values suggest a direct correlation, where higher values of the feature are associated with higher obesity rates; whereas negative values indicate an inverse relationship, suggesting that higher feature values are associated with lower obesity rates. For example, Feature 1112th shows a strong negative correlation of −0.6061, implying that higher values of this feature are associated with lower obesity rates.

In Figure 8, Feature 1112th is identified as the most significant predictor of obesity rates among the 2048 features, marking its top rank in importance. A comparative analysis of the heat maps for Feature 1112th and obesity rates reveals a strong correlation between the two, as evidenced by Table 2 data where Feature 1112th has an importance of 15.01% and a predicted obesity correlation coefficient of −0.6061. This negative correlation suggests that higher values of Feature 1112th are associated with lower obesity rates. Feature 1112th exhibits high values in densely populated urban areas and frequently visited tourist destinations. The cities of Kansas City (population: 510,704), St. Louis (population: 281,754), Springfield (population: 170,188), and Columbia (population: 129,330) are the top four most populated cities in Missouri, according to the 2023 Population Estimates Program and the 2022 American Community Survey. All these cities are prominently highlighted on the heat map for Feature 1112th, aligning with their lower obesity rates due to the negative correlation indicated in the Table 2. Interestingly, the area around Branson, Missouri, despite having a relatively low full-time resident population of 12,897, also shows high values for Feature 1112th. Branson's high feature values can be attributed to its status as a major tourist destination, which attracts millions of visitors annually. This suggests that Feature 1112th captures geospatial aspects related to urban development and human activity, such as:

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

13 of 23

1. High Building Density and Commercial Structures: Frequent clustering of buildings and commercial areas.
2. Extensive Transportation Networks: Well-developed roads, highways, and parking facilities.
3. Presence of Recreational Areas: Parks, theaters, and other recreational facilities indicating significant human activity.
4. Impervious Surfaces: High proportion of concrete and asphalt surfaces.
5. Shadow Patterns: Distinct patterns cast by tall structures.
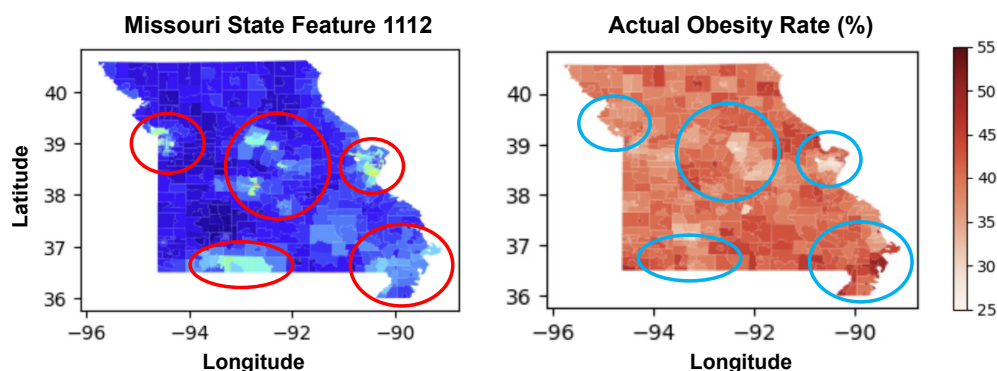
Figure 8. The left map shows spatial distribution of Feature 1112 across Missouri, with red circles highlighting areas of high values (urban areas). The right map depicts actual obesity rates (%) across the state, with blue circles indicating regions with lower obesity prevalence. Notable discrepancies between feature values and obesity rates can be observed in several regions.

The correlation between the heat maps of Feature 1112th and obesity rates, along with the feature's association with densely populated and highly visited areas, indicates that Feature 1112th likely reflects critical geospatial aspects of urban environments. These aspects include building density, transportation networks, commercial and recreational infrastructure, and other indicators of human activity and development. Further research is necessary to decompose this feature and understand the specific attributes it encompasses, which could inform targeted public health interventions and urban planning policies.

4.4. Additional Top-5 Features

Figure 9, illustrates the geospatial distribution of two significant features across Missouri, identified as Feature 0095th and Feature 1314th, which are ranked as the second and third most important features relative to obesity rates. It is noteworthy that higher values of these features, particularly concentrated in urban centers such as Kansas City and St. Louis, are associated with lower obesity rates. This suggests that these features, which could represent factors such as access to recreational facilities or health services, play a crucial role in mitigating obesity rates in these areas.

Figure 10, presents two heat maps illustrating the spatial distribution of important features across Missouri, identified as Feature 0767th and Feature 0239th. Higher values of these features are associated with lower obesity rates across Missouri's various regions. The visual patterns suggest areas where targeted interventions might be beneficial in reducing obesity prevalence.

Table 3 provides a numerical summary of the performance metrics for a 10-fold cross-validation of a regression model predicting obesity rates. Each row in the table corresponds to one of the ten folds used in the cross-validation process. The columns present the MSE, R^2, and Adjusted R^2 for each fold. The MSE is a measure of the average squared difference between the observed actual outcomes and the outcomes predicted by the model. The R^2 value indicates the proportion of the variance for the dependent variable that's explained by the independent variables in the model, while the adjusted R^2 accounts for the number of predictors in the model and provides a more adjusted measure of the goodness of fit.
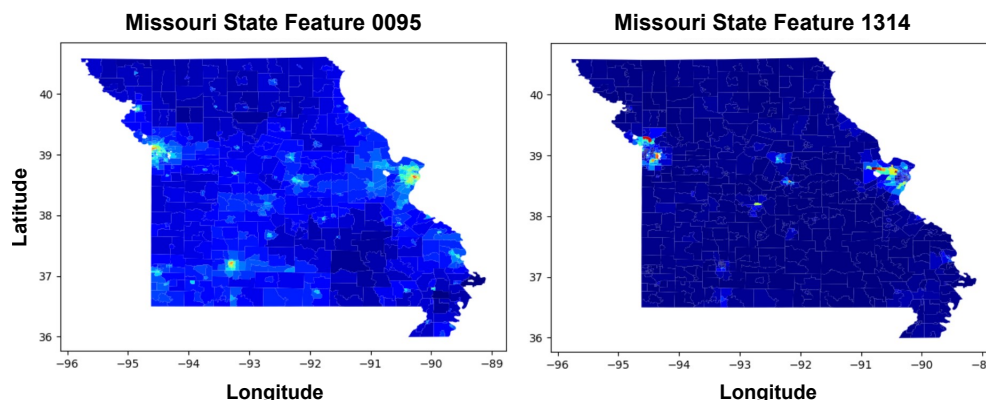
Int. J. Environ. Res. Public Health **2024**, *21*, 1534

14 of 23

Figure 9. The spatial distribution of the 2nd and 3rd most important features across Missouri, (**left**) Feature 0095th and (**right**) Feature 1314. Urban areas, particularly around Kansas City for Feature 0095 and St. Louis for Feature 1314, show significant concentrations of these features.
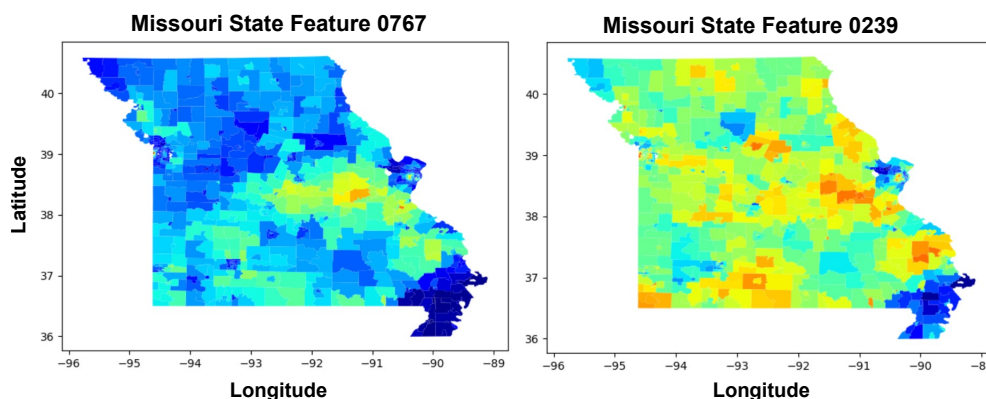

Figure 10. Spatial distribution of the 4th and 5th most important features across Missouri, (**left**) Feature 0767 and (**right**) Feature 0239.

Table 3. Provides a numerical summary of the performance metrics for a 10-fold cross-validation of a regression model predicting obesity rates.

Fold #	MSE	R^2	Adjusted R^2
1	18.54	0.406	0.400
2	22.06	0.011	0.001
3	12.56	0.538	0.534
4	13.07	0.546	0.541
5	13.58	0.559	0.555
6	10.22	0.661	0.657
7	12.81	0.537	0.533
8	16.68	0.485	0.480
9	11.06	0.567	0.563
10	16.36	0.456	0.451
Mean	14.69	0.477	0.471

4.5. DCNN Model Explainability

Figure 11 showcases four distinct heatmaps analyzing obesity rates and prediction accuracy across the state of Missouri. Figure 11A is a heatmap which displays the actual obesity rates, where color intensities represent the percentage of the population affected by obesity, ranging from 25% to 55%. Darker shades correspond to areas with higher obesity rates, allowing for the identification of regions where obesity is more prevalent.
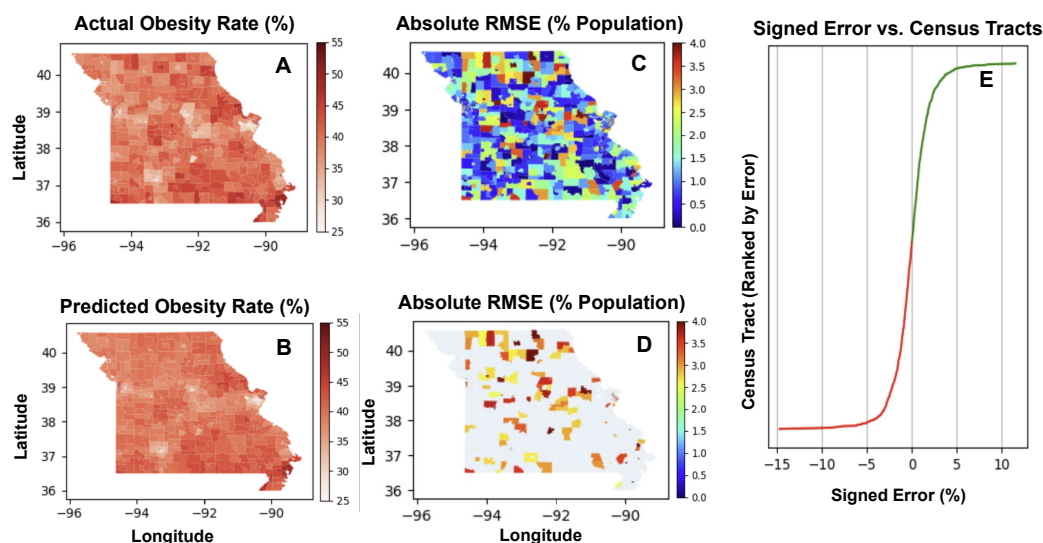
Int. J. Environ. Res. Public Health **2024**, *21*, 1534

15 of 23

Figure 11. Geospatial analysis of obesity rates and prediction accuracy in Missouri. (**A**) displays the actual obesity rates, while (**B**) shows the predicted rates, both using a color gradient to represent percentages. (**C**) highlights areas with significant predictive errors by filtering out RMSE values below 4, focusing on regions where the model's accuracy is lower. (**D**) refines this analysis by presenting a broader error distribution, including RMSE values of 2.5 and above, using the same color gradient for consistency. (**E**) Curve illustrating the signed error distribution of predicted obesity rates across census tracts. Negative signed errors indicate underpredictions (shown in red) and positive errors indicate overpredictions (shown in green). The census tracts are ranked by the magnitude of error, highlighting the asymmetry in predictive accuracy and potential systematic bias in the model.

Figure 11B illustrates the obesity rates as predicted by a statistical model. The visual similarity between Figure 11A,B suggests a close alignment between predicted and actual rates, indicating the model's effectiveness in mirroring the geographical distribution of obesity.

Figure 11C applies a filter to the absolute RMSE values, excluding any areas where the RMSE is lower than 4. This effectively removes more accurately predicted regions, highlighting only those areas where the predictive model's errors are above this threshold. The resulting map is a patchwork of colors where only the highest errors in prediction are shown, offering a focused view on the model's limitations.

Figure 11D refines the approach in Figure 11C by adjusting the RMSE filter to include areas with RMSE values of 2.5 and above. The same color ratio is maintained to ensure consistency in visual interpretation across Figure 11C,D. As a result, this figure shows a broader range of prediction errors compared to Figure 11C, providing a more nuanced understanding of the model's performance across the region.

The progression from Figure 11A,B to Figure 11C,D demonstrates a methodical approach to analyzing model accuracy. The initial heatmaps set the stage by presenting actual and predicted rates, while the latter heatmaps critically evaluate the model by highlighting regions with the most significant prediction errors. This step-by-step visualization emphasizes areas where the predictive model could be improved, guiding researchers and policymakers in directing their efforts to refine predictive analytics and address obesity more effectively in regions with higher prediction errors.

Figure 11E features a curve that represents the signed error between predicted and actual values across different census tracts. The horizontal axis is labeled "Signed Error (%)" and indicates the percentage error of prediction, with negative values on the left side representing under-predictions and positive values on the right side indicating over-predictions. The vertical axis is labeled "Census Tract Ranked by Error", suggesting that each census tract has been ranked based on the magnitude of the prediction error.

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

16 of 23

The curve starts on the left with a steep incline in the red zone, indicating a significant under-prediction of obesity rates in some census tracts (those with a high negative signed error). As the curve moves to the right, the color transitions to green, and the slope decreases until the error approaches zero. The zero point on the horizontal axis likely represents the point at which the prediction exactly matches the actual obesity rate.

Beyond the zero point, the curve extends into the "Over Predict" zone, where the signed error becomes positive, indicating census tracts where the model has overestimated the obesity rate. The curve in this region is relatively flat and continues in green, suggesting fewer tracts with over-prediction compared to under-prediction.

The shape of the curve indicates that the distribution of errors is skewed; there are more census tracts with substantial under-predictions than over-predictions. This could imply that the model used for prediction has a systematic bias or that certain factors leading to higher obesity rates are not being adequately captured in those tracts with higher negative errors.

Table 4 is the data extracted from the larger dataset, and is related to our study on obesity prevalence within various census tracts. Each row corresponds to a specific census tract as indicated by the "GEOID" column, which contains unique numerical identifiers. The "County" column lists the location of each census tract, with multiple entries for the urban counties of St. Louis and Jackson, and only one entry for the rural counties of Stoddard and Pemiscot.

Table 4. The table lists the GEOID, County, Population, and actual versus predicted obesity percentages for selected census tracts, along with the signed error indicating the discrepancy between predicted and observed values.

GEOID	County	Population	Actual Obesity (%)	Predicted Obesity (%)	Signed Error (%)
29510124600	St. Louis	1845	46.00	37.63	8.37
29207470600	Stoddard	4968	39.80	48.32	−8.52
29095005100	Jackson	1651	34.00	42.76	−8.76
29095016100	Jackson	2046	49.80	40.42	9.38
29095008200	Jackson	2765	30.50	40.14	−9.64
29510119300	St. Louis	5454	26.90	37.82	−10.92
29095015400	Jackson	3826	52.80	41.61	11.19
29155470200	Pemiscot	3613	50.80	39.31	11.49
29095006100	Jackson	2542	48.30	61.83	−13.53
29189213400	St. Louis	6669	37.40	52.07	−14.67

The "Population" column lists the number of people residing in each census tract. These populations range from as low as 1651 to as high as 6669, suggesting a diverse set of tracts in terms of population size.

In the "Actual Obesity (%)" column, we see the actual percentage of the population within each tract that has been classified as obese. These percentages range from 26.90% to 52.80%, indicating a significant variation in obesity rates across different areas.

The "Predicted Obesity (%)" column shows the predicted obesity rates as for the same census tracts. The predicted percentages exhibit a similar range to the actual percentages but do not always align exactly, which is to be expected in predictive modeling.

The final column, "Signed Error (%)", represents the difference between the predicted and actual obesity rates, expressed as a percentage. A positive signed error indicates that the model has over-predicted the obesity rate, while a negative signed error suggests under-prediction. For instance, the first row shows an over-prediction of 8.37% for the census tract with GEOID 29510124600, meaning the model estimated the obesity rate to be 8.37 percentage points lower than the actual rate. Conversely, the second row, with GEOID 29207470600, shows an under-prediction of 8.52%, where the model's estimate was higher than the actual rate.

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

17 of 23

5. Discussion

The visual contrasts in Figure 3 underscore the presence of geographical health disparities that may warrant targeted public health interventions and resource allocation. However, the map also highlights the challenge of representing densely populated areas where census tracts are smaller, and thus, the corresponding data may be less discernible. In all, the study shows that in both over-sampled (large) census tracts and under-sampled (very small) census tracts the model performs well.

The MSE for each fold is a measure of the average squared difference between the observed actual outcomes and the outcomes predicted by the model. The values range from a low of 10.22 in Fold 6 to a high of 22.06 in Fold 2, indicating variability in the model's predictive accuracy across different subsets of data [73–75]. This is a common phenomena referred to as finite sample effects [76], which indicates that with limited data, the specific examples in each fold can have a disproportionate influence on model training. Some folds might, by chance, contain more challenging or unusual examples.

The R^2 values range from as low as 0.011 in Fold 2 to as high as 0.661 in Fold 6. The R^2 metric provides a measure of how well the variation in the dependent variable (obesity rates) is explained by the model. In this context, Fold 6's model explains 66.1% of the variance, which is quite high, whereas the model in Fold 2 explains only 1.1%, which is extremely low.

Adjusted R^2 is a modified version of R^2 [77] that adjusts for the number of predictors in the model. It is generally lower than R^2 as it penalizes excessive use of predictors. The Adjusted R^2 values closely follow the R^2 values, with the highest being 0.657 in Fold 6 and the lowest being 0.001 in Fold 2.

The last row provides the mean of the MSE, R^2, and Adjusted R^2 across all folds. The mean MSE of 14.69 suggests that on average, the model has a moderate level of prediction error. The mean R^2 of 0.477 indicates that, on average, the model explains about 47.7% of the variability in the actual data, which suggests a moderate fit. Similarly, the mean Adjusted R^2 of 0.471 is in line with the mean R^2, indicating a reasonable fit after adjusting for the number of predictors.

Table 3 highlights that the performance of the model varies significantly across the different folds, as evidenced by the range of values in MSE and R^2 metrics. This variation can arise due to the inherent differences in the data subsets or may indicate an underlying issue with the model's stability. Table 3 underscores the importance of cross-validation to understand how the model might perform in general, rather than relying on the metrics from a single partition of the data. Overall, Table 3 provides a detailed breakdown of the model's performance across different segments of the data, highlighting variations in model accuracy and fit.

Figure 11E provides a visual representation of the prediction errors across census tracts, with the curve's profile offering insight into the model's predictive performance and potential biases. The clear demarcation of under-prediction and over-prediction zones also helps in identifying the direction of the errors for subsequent analysis and model refinement.

Table 4 is likely used to analyze the accuracy of the obesity rate predictions by the model, providing insight into where the model tends to overestimate or underestimate obesity prevalence. This information could be valuable for refining the model or for directing public health resources to areas where obesity is more prevalent than initially predicted.

Despite the promising results, our study has some limitations. First, the estimates of obesity prevalence from the Behavioral Risk Factor Surveillance System rely on self-reported measurements of height and weight, which are subject to bias and often result in an underestimation of the true rate of obesity [37,39]. Variations in the timing between when the obesity data and the satellite images are collected can also introduce biases into our analysis. One of the primary limitations of our study pertains to the dataset's size and geographical coverage. The research was confined to 1052 census tracts within the state of Missouri, limiting the generalizability of the findings.

Int. J. Environ. Res. Public Health **2024**, 21, 1534

18 of 23

Although these tracts were selected to represent a diverse range of urban and rural areas, they do not encompass the neighboring states and their varied demographic and geographic profiles. Furthermore, the limited number of census tracts might not provide a sufficiently robust dataset for more complex machine learning models [78]. The pre-trained ResNet-50 [59] is not fully optimized for the specific nuances of satellite image analysis related to obesity rate prediction. Our study's findings must therefore be interpreted with caution, acknowledging that the employed models, although advanced, might not capture the complete range of factors influencing obesity rates as discernible from satellite imagery. Additionally, training the model to adjust its weights to satellite data and the chosen census tracts should yield significantly better results.

6. Conclusions

In summary, the main findings of our research study are to emphasize the novel methodology using DNVF to analyze and predict the obesity prevalence for different census tracts in Missouri. In conclusion, the methodology employed in this research holds great promise for enhancing our understanding and predicting obesity. It has the potential to inform more effective community-level and policy-driven strategies to combat this complex health issue.

Although the study is innovative in its approach. The reliance on self-reported data could introduce biases, and the focus solely on Missouri limits the generalizability of the findings. These aspects, however, present valuable opportunities for future research. In the future, it will be important to expand the geographical scope of this study to include diverse regions, enhancing the generalizability of our findings. We also plan to integrate additional data sources, like socio-economic and health-related factors, to provide a more comprehensive analysis. The refinement of machine learning techniques, particularly advanced models and deep learning, will be pivotal in improving prediction accuracy. Longitudinal and geospatial studies are envisioned to observe temporal changes in obesity rates and other population health diseases.

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

19 of 23

Abbreviations

The following abbreviations are used in this manuscript:

DCNN   Deep Convolutional Neural Networks
DNVF   Deep Neural Visual Features
MSE   Mean Squared Error
BRFSS   Behavioral Risk Factor Surveillance System
BMI   Body Mass Index
CDC   Centers for Disease Control and Prevention
CT   Census Tract
ML   Machine Learning
DL   Deep Learning

Appendix A

In Figure A1, feature 1112th is the most significant, contributing over 15% to the model's outcomes. The other features vary in importance from just under 3.5% to approximately 1%, with feature numbers 0095th, 1314th, 0767th, 0239th, 1253th, 0895th, 1126th, 0338th, and 0668th following in descending order of importance. The bar chart highlights the dominant influence of feature 1112th compared to the other visual features.
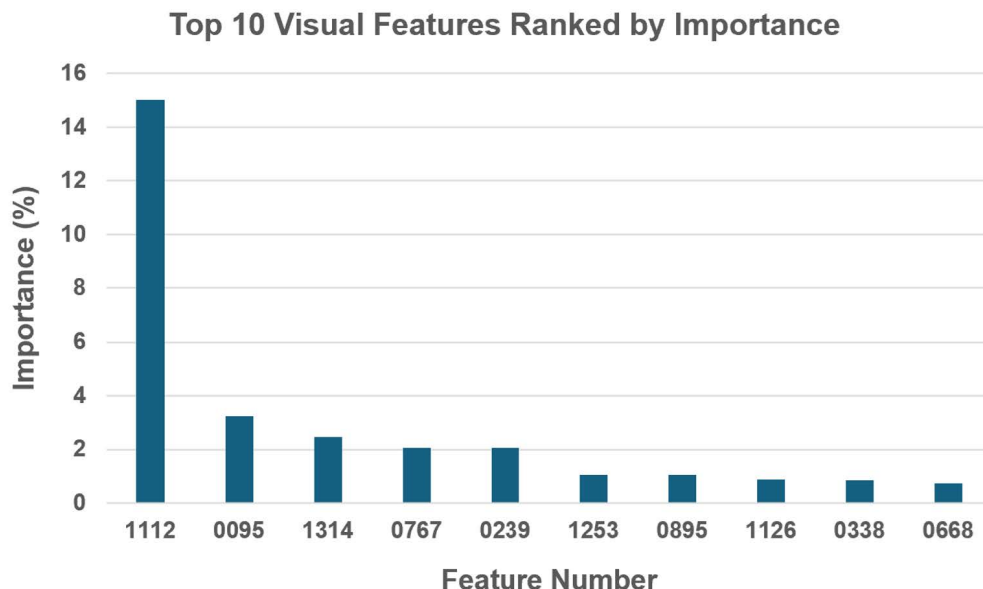

Figure A1. Bar chart displays the top 10 visual features ranked by their importance as used in our obesity rates prediction model. The x-axis represents the feature numbers, which are specific identifiers for each visual feature. The y-axis indicates the importance of each feature in percentage terms.

Appendix B

Table A1 provides a breakdown of the top ten features in terms of their importance percentages, derived from the 10-fold-cross validation model aiming to predict the obesity rates using Deep Neural Visual Features (DNVFs). Feature 1112th is the most significant, holding an importance of 15.01%, indicating its strong influence on the model's predictions regarding obesity rates. The subsequent features, although less important than Feature 1112th, still contribute notably to the model's prediction and include Features 0095th and 1314th, with importance percentages of 3.23% and 2.47%, respectively.

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

20 of 23

Table A1. This table lists the top ten visual features ranked by their importance in our obesity rate prediction model. The percentages indicate the relative importance of each feature in influencing the model's predictions, with higher values signifying greater influence. The 1112th is the most important, with a corresponding importance of 15.01%.

Feature #	Importance (%)
1112	15.01
0095	3.23
1314	2.47
0767	2.07
0239	2.05
1253	1.05
0895	1.04
1126	0.87
0338	0.84
0668	0.75

References

1. Sung, B.; Etemadifar, A. Multilevel analysis of socio-demographic disparities in adulthood obesity across the United States geographic regions. *Osong Public Health Res. Perspect.* **2019**, *10*, 137. [CrossRef] [PubMed]
2. Maharana, A.; Nsoesie, E.O. Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity. *JAMA Netw. Open* **2018**, *1*, e181535. [CrossRef] [PubMed]
3. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
4. Jean, N.; Burke, M.; Xie, M.; Davis, W.M.; Lobell, D.B.; Ermon, S. Combining satellite imagery and machine learning to predict poverty. *Science* **2016**, *353*, 790–794. [CrossRef]
5. Almalaysha, M.; Singh, A.; Muhsin, S.A.; Morey, A.; Zhang, S.; Channaiah, L.H.; Almasri, M. Microfluidic Biosensor for Rapid Detection of Salmonella in Raw Chicken Products. In Proceedings of the 2024 IEEE 37th International Conference on Micro Electro Mechanical Systems (MEMS), Austin, TX, USA, 21–25 January 2024; pp. 308–311.
6. Almalaysha, M.; Singh, A.; Muhsin, S.A.; Carlson, A.V.; Trout, K.E.; Morey, A.; Zhang, S.; Channaiah, L.H.; Almasri, M. A highly sensitive microfluidic biosensor for rapid and accurate detection of salmonella in raw chicken products. *Sensors Actuators Rep.* **2024**, *9*, 100257. [CrossRef]
7. Dahu, M.; Khan, S.; Toubal, I.E.; Alshehri, M.; Martinez-Villar, C.I.; Ogundele, O.B.; Sheets, L.R.; Scott, G.J. Geospatial Modeling of Deep Neural Visual Features for Predicting Obesity Prevalence in Missouri. *JMIR Preprints* **2024**, *3*, 64362. [CrossRef]
8. Marmanis, D.; Datcu, M.; Esch, T.; Stilla, U. Deep learning earth observation classification using ImageNet pretrained networks. *IEEE Geosci. Remote. Sens. Lett.* **2015**, *13*, 105–109. [CrossRef]
9. Dahu, B.M.; Alaboud, K.; Nowbuth, A.A.; Puckett, H.M.; Scott, G.J.; Sheets, L.R. The Role of Remote Sensing and Geospatial Analysis for Understanding COVID-19 Population Severity: A Systematic Review. *Int. J. Environ. Res. Public Health* **2023**, *20*, 4298. [CrossRef] [PubMed]
10. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
11. Cao, Y.; Dahu, B.M.; Scott, G.J. A geographic computational visual feature database for natural and anthropogenic phenomena analysis from multi-resolution remote sensing imagery. In Proceedings of the ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, Seattle, WA, USA, 3 November 2020; pp. 1–10.
12. Alaboud, K.; Toubal, I.E.; Dahu, B.M.; Daken, A.A.; Salman, A.A.; Alaji, N.; Hamadeh, W.; Aburayya, A. The Quality Application of Deep Learning in Clinical Outcome Predictions Using Electronic Health Record Data: A Systematic Review. *South East. Eur. J. Public Health* **2023**, *XXI*, 9–23.
13. Nguyen, Q.C.; Keralis, J.M.; Dwivedi, P.; Ng, A.E.; Javanmardi, M.; Khanna, S.; Huang, Y.; Brunisholz, K.D.; Kumar, A.; Tasdizen, T. Leveraging 31 million Google street view images to characterize built environments and examine county health outcomes. *Public Health Rep.* **2021**, *136*, 201–211. [CrossRef] [PubMed]
14. Ohanyan, H.; Portengen, L.; Huss, A.; Traini, E.; Beulens, J.W.; Hoek, G.; Lakerveld, J.; Vermeulen, R. Machine learning approaches to characterize the obesogenic urban exposome. *Environ. Int.* **2022**, *158*, 107015. [CrossRef] [PubMed]
15. Hong, X.; Sheridan, S.; Li, D. Mapping built environments from UAV imagery: A tutorial on mixed methods of deep learning and GIS. *Comput. Urban Sci.* **2022**, *2*, 12. [CrossRef]
16. Larkin, A.; Krishna, A.; Chen, L.; Amram, O.; Avery, A.R.; Duncan, G.E.; Hystad, P. Measuring and modelling perceptions of the built environment for epidemiological research using crowd-sourcing and image-based deep learning models. *J. Expo. Sci. Environ. Epidemiol.* **2022**, *32*, 892–899. [CrossRef] [PubMed]

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

21 of 23

17. Wang, R.; Liu, Y.; Lu, Y.; Zhang, J.; Liu, P.; Yao, Y.; Grekousis, G. Perceptions of built environment and health outcomes for older Chinese in Beijing: A big data approach with street view images and deep learning technique. *Comput. Environ. Urban Syst.* **2019**, *78*, 101386. [CrossRef]

18. Sallis, J.F.; Floyd, M.F.; Rodríguez, D.A.; Saelens, B.E. Role of built environments in physical activity, obesity, and cardiovascular disease. *Circulation* **2012**, *125*, 729–737. [CrossRef]

19. Zhou, H.; He, S.; Cai, Y.; Wang, M.; Su, S. Social inequalities in neighborhood visual walkability: Using street view imagery and deep learning technologies to facilitate healthy city planning. *Sustain. Cities Soc.* **2019**, *50*, 101605. [CrossRef]

20. Alshehri, M.; Ouadou, A.; Scott, G.J. Deep Transformer-based Network Deforestation Detection in the Brazilian Amazon Using Sentinel-2 Imagery. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 2502705. [CrossRef]

21. Rachele, J.N.; Wang, J.; Wijnands, J.S.; Zhao, H.; Bentley, R.; Stevenson, M. Using machine learning to examine associations between the built environment and physical function: A feasibility study. *Health Place* **2021**, *70*, 102601. [CrossRef]

22. Giles-Corti, B.; Macintyre, S.; Clarkson, J.P.; Pikora, T.; Donovan, R.J. Environmental and lifestyle factors associated with overweight and obesity in Perth, Australia. *Am. J. Health Promot.* **2003**, *18*, 93–102. [CrossRef]

23. Saelens, B.E.; Sallis, J.F.; Black, J.B.; Chen, D. Neighborhood-based differences in physical activity: An environment scale evaluation. *Am. J. Public Health* **2003**, *93*, 1552–1558. [CrossRef] [PubMed]

24. Burdette, H.L.; Whitaker, R.C. Neighborhood playgrounds, fast food restaurants, and crime: Relationships to overweight in low-income preschool children. *Prev. Med.* **2004**, *38*, 57–63. [CrossRef]

25. Black, J.L.; Macinko, J. Neighborhoods and obesity. *Nutr. Rev.* **2008**, *66*, 2–20. [CrossRef]

26. Inagami, S.; Cohen, D.A.; Finch, B.K.; Asch, S.M. You are where you shop: Grocery store locations, weight, and neighborhoods. *Am. J. Prev. Med.* **2006**, *31*, 10–17. [CrossRef]

27. Rundle, A.; Roux, A.V.D.; Freeman, L.M.; Miller, D.; Neckerman, K.M.; Weiss, C.C. The urban built environment and obesity in New York City: A multilevel analysis. *Am. J. Health Promot.* **2007**, *21*, 326–334. [CrossRef]

28. Khan, S.; Dahu, B.M.; Scott, G.J. A Spatio-temporal Study of Changes in Air Quality from Pre-COVID Era to Post-COVID Era in Chicago, USA. *Aerosol Air Qual. Res.* **2022**, *22*, 220053. [CrossRef]

29. Rutt, C.D.; Coleman, K.J. Examining the relationships among built environment, physical activity, and body mass index in El Paso, TX. *Prev. Med.* **2005**, *40*, 831–841. [CrossRef] [PubMed]

30. Li, F.; Harmer, P.; Cardinal, B.J.; Bosworth, M.; Johnson-Shelton, D.; Moore, J.M.; Acock, A.; Vongjaturapat, N. Built environment and 1-year change in weight and waist circumference in middle-aged and older adults: Portland Neighborhood Environment and Health Study. *Am. J. Epidemiol.* **2009**, *169*, 401–408. [CrossRef]

31. Dahu, B.M.; Khan, S.; Salman, A.A.; Andraws, Y.M.; Daken, A.A.; Aburayya, A. Epidemiological Analysis of Vaccination Strategies and Demographic Patterns In COVID-19 Cases in the Midwest Region of the United States. *Natl. J. Community Med.* **2024**, *15*, 62–71. [CrossRef]

32. Rundle, A.; Neckerman, K.M.; Freeman, L.; Lovasi, G.S.; Purciel, M.; Quinn, J.; Richards, C.; Sircar, N.; Weiss, C. Neighborhood food environment and walkability predict obesity in New York City. *Environ. Health Perspect.* **2009**, *117*, 442–447. [CrossRef]

33. Li, F.; Harmer, P.A.; Cardinal, B.J.; Bosworth, M.; Acock, A.; Johnson-Shelton, D.; Moore, J.M. Built environment, adiposity, and physical activity in adults aged 50–75. *Am. J. Prev. Med.* **2008**, *35*, 38–46. [CrossRef] [PubMed]

34. Dahu, B.M.; Khan, S.; Li, W.S.; Shu, X.; Woldu, H.; Popescu, M.; Sheets, L.R.; Scott, G.J. Demographic and Time Trend Analysis of COVID-19 Test Results of Boone County, Missouri. *AMIA Summits Transl. Sci. Proc.* **2023**, *2023*, 91. [PubMed]

35. Jeffery, R.; Baxter, J.; McGuire, M.; Linde, J. Are fast food restaurants an environmental risk factor for obesity? *Int. J. Behav. Nutr. Phys. Act.* **2006**, *3*, 2. [CrossRef] [PubMed]

36. Lovasi, G.S.; Hutson, M.A.; Guerra, M.; Neckerman, K.M. Built environments and obesity in disadvantaged populations. *Epidemiol. Rev.* **2009**, *31*, 7–20. [CrossRef]

37. Kuczmarski, M.F.; Kuczmarski, R.J.; Najjar, M. Effects of age on validity of self-reported height, weight, and body mass index: Findings from the Third National Health and Nutrition Examination Survey, 1988–1994. *J. Am. Diet. Assoc.* **2001**, *101*, 28–34. [CrossRef] [PubMed]

38. Papas, M.A.; Alberg, A.J.; Ewing, R.; Helzlsouer, K.J.; Gary, T.L.; Klassen, A.C. The built environment and obesity. *Epidemiol. Rev.* **2007**, *29*, 129–143. [CrossRef]

39. Merrill, R.M.; Richardson, J.S. Peer reviewed: Validity of self-reported height, weight, and body mass index: Findings from the National Health and nutrition examination survey, 2001–2006. *Prev. Chronic Dis.* **2009**, *6*, A121.

40. Dahu, B.M. Demographic and Trend Analysis of COVID-19 Test Results of Boone County, Missouri. Ph.D. Thesis, University of Missouri-Columbia, Columbia, MO, USA, 2022.

41. Zheng, W.; McLerran, D.F.; Rolland, B.; Zhang, X.; Inoue, M.; Matsuo, K.; He, J.; Gupta, P.C.; Ramadas, K.; Tsugane, S.; et al. Association between body-mass index and risk of death in more than 1 million Asians. *N. Engl. J. Med.* **2011**, *364*, 719–729. [CrossRef]

42. Jafar, T.H.; Islam, M.; Poulter, N.; Hatcher, J.; Schmid, C.H.; Levey, A.S.; Chaturvedi, N. Children in south Asia have higher body mass–adjusted blood pressure levels than white children in the United States: A comparative study. *Circulation* **2005**, *111*, 1291–1297. [CrossRef]

43. Hales, C.M.; Carroll, M.D.; Fryar, C.D.; Ogden, C.L. *Prevalence of Obesity Among Adults and Youth: United States, 2015–2016*; NCHS Data Brief No. 288; Centers for Disease Control and Prevention: Atlanta, GA, USA, 2017.

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

22 of 23

44. Casagrande, S.S.; Whitt-Glover, M.C.; Lancaster, K.J.; Odoms-Young, A.M.; Gary, T.L. Built environment and health behaviors among African Americans: A systematic review. *Am. J. Prev. Med.* **2009**, *36*, 174–181. [CrossRef]

45. Dahu, B.M.; Aburayya, A.; Shameem, B.; Shwedeh, F.; Alawadhi, M.; Aljasmi, S.; Salloum, S.A.; Aburayya, H.; Aburayya, I. The impact of COVID-19 lockdowns on air quality: A systematic review study. *South East. Eur. J. Public Health* **2022**, *5*, 1–22. [CrossRef]

46. Nguyen, Q.C.; Sajjadi, M.; McCullough, M.; Pham, M.; Nguyen, T.T.; Yu, W.; Meng, H.W.; Wen, M.; Li, F.; Smith, K.R.; et al. Neighbourhood looking glass: 360° automated characterisation of the built environment for neighbourhood effects research. *J. Epidemiol. Community Health* **2018**, *72*, 260–266. [CrossRef] [PubMed]

47. Songhyeon, S.; Jahun, K.; Sangho, C. A Study on the Effect of Urban Environment on Obesity Using Multinomial Logistic Regression Models. *J. Korean Soc. Transp.* **2022**, *40*, 717–734.

48. Lam, T.M.; Vaartjes, I.; Grobbee, D.E.; Karssenberg, D.; Lakerveld, J. Associations between the built environment and obesity: An umbrella review. *Int. J. Health Geogr.* **2021**, *20*, 7. [CrossRef] [PubMed]

49. Alkhalaf, M.; Yu, P.; Shen, J.; Deng, C. A review of the application of machine learning in adult obesity studies. *Appl. Comput. Intell.* **2022**, *2*, 32–48. [CrossRef]

50. An, R.; Ji, M.; Yan, H.; Guan, C. Impact of ambient air pollution on obesity: A systematic review. *Int. J. Obes.* **2018**, *42*, 1112–1126. [CrossRef]

51. Zhou, X.; Chen, L.; Liu, H.X. Applications of Machine Learning Models to Predict and Prevent Obesity: A Mini-Review. *Front. Nutr.* **2022**, *9*, 933130. [CrossRef]

52. Toubal, I.E.; Lyu, L.; Lin, D.; Palaniappan, K. Single view facial age estimation using deep learning with cascaded random forests. In Proceedings of the Computer Analysis of Images and Patterns: 19th International Conference (CAIP 2021), Virtual Event, 28–30 September 2021; Springer: Cham, Switzerland, 2021; pp. 285–296.

53. Toubal, I.E.; Duan, Y.; Yang, D. Deep learning semantic segmentation for high-resolution medical volumes. In Proceedings of the 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, USA, 13–15 October 2020; pp. 1–9.

54. Toubal, I.E.; Al-Shakarji, N.; Cornelison, D.; Palaniappan, K. Ensemble deep learning object detection fusion for cell tracking, mitosis, and lineage. *IEEE Open J. Eng. Med. Biol.* **2023**. [CrossRef]

55. Newton, D.; Piatkowski, D.; Marshall, W.; Tendle, A. Deep Learning Methods for Urban Analysis and Health Estimation of Obesity. In Proceedings of the 38th ECAADe Conference on Anthropologic: Architecture and Fabrication in the Cognitive Age, Berlin, Germany, 16–17 September 2020; Volume 1, pp. 297–304.

56. Yue, X.; Antonietti, A.; Alirezaei, M.; Tasdizen, T.; Li, D.; Nguyen, L.; Mane, H.; Sun, A.; Hu, M.; Whitaker, R.T.; et al. Using convolutional neural networks to derive neighborhood built environments from Google Street View images and examine their associations with health outcomes. *Int. J. Environ. Res. Public Health* **2022**, *19*, 12095. [CrossRef]

57. Phan, L.; Yu, W.; Keralis, J.M.; Mukhija, K.; Dwivedi, P.; Brunisholz, K.D.; Javanmardi, M.; Tasdizen, T.; Nguyen, Q.C. Google street view derived built environment indicators and associations with state-level obesity, physical activity, and chronic disease mortality in the united states. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3659. [CrossRef]

58. Simonyan, K. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

59. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

60. United States Census Bureau. TIGER/Line Shapefiles. 2022. Available online: https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2022.html (accessed on 14 September 2023).

61. Centers for Disease Control and Prevention. Adult Obesity Facts. 2022. Available online: https://www.cdc.gov/obesity/data/adult.html (accessed on 4 September 2023).

62. Alshehri, M.; Ouadou, A.; Scott, G. Deforestation Detection in the Brazilian Amazon Using Transformer-based Networks. In Proceedings of the 2023 IEEE Conference on Artificial Intelligence (CAI), Santa Clara, CA, USA, 5–6 June 2023; pp. 292–293.

63. Dahu, B.M.; Khan, S.; Sheets, L.R.; Scott, G.J. Exploring the Geospatial Relationship Between COVID-19 Positivity and Income in Mixed Urban-Rural Population. *Stud. Health Technol. Inform.* **2024**, *310*, 1327–1331. [PubMed]

64. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]

65. Mascarenhas, S.; Agarwal, M. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In Proceedings of the 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), Bengaluru, India, 19–21 November 2021; Volume 1, pp. 96–99.

66. Rathi, P.; Kuwar Gupta, R.; Agarwal, S.; Shukla, A. Sign language recognition using resnet50 deep neural network architecture. In Proceedings of the 5th International Conference on Next Generation Computing Technologies (NGCT-2019), Dehradun, India, 20–21 December 2019.

67. Malakouti, S.M.; Menhaj, M.B.; Suratgar, A.A. The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction. *Clean. Eng. Technol.* **2023**, *15*, 100664. [CrossRef]

68. Dor, O.; Zhou, Y. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins Struct. Funct. Bioinform.* **2007**, *66*, 838–845. [CrossRef] [PubMed]

Int. J. Environ. Res. Public Health **2024**, *21*, 1534

23 of 23

69. Nowbuth, A.A.; Asombang, A.W.; Alaboud, K.; Souque, C.; Dahu, B.M.; Pather, K.; Mwanza, M.M.; Lotfi, S.; Parmar, V.S. Gamification as an educational tool to address antimicrobial resistance: A systematic review. *JAC-Antimicrob. Resist.* **2023**, *5*, dlad130. [CrossRef]

70. Potts, S.E.; Rose, K.A. Evaluation of GLM and GAM for estimating population indices from fishery independent surveys. *Fish. Res.* **2018**, *208*, 167–178. [CrossRef]

71. Gauthier, P.A.; Scullion, W.; Berry, A. Sound quality prediction based on systematic metric selection and shrinkage: Comparison of stepwise, lasso, and elastic-net algorithms and clustering preprocessing. *J. Sound Vib.* **2017**, *400*, 134–153. [CrossRef]

72. Giannakas, F.; Troussas, C.; Krouska, A.; Sgouropoulou, C.; Voyiatzis, I. Xgboost and deep neural network comparison: The case of teams' performance. In Proceedings of the 17th International Conference on Intelligent Tutoring Systems (ITS 2021), Virtual Event, 7–11 June 2021; Springer: Cham, Switzerland, 2021; pp. 343–349.

73. Bengio, Y.; Grandvalet, Y. No unbiased estimator of the variance of k-fold cross-validation. *Adv. Neural Inf. Process. Syst.* **2003**, *16*, 1–12.

74. Burman, P. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika* **1989**, *76*, 503–514. [CrossRef]

75. Nowbuth, A.; Asombang, A.; Alaboud, K.; Souque, C.; Dahu, B.; Pather, K.; Mwanza, M.; Lotfi, S.; Parmar, V. Evaluating Gamification of Antimicrobial Resistance as an Educational Tool: A Systematic Review. *Eur. J. Public Health* **2023**, *33*, ckad160.1013. [CrossRef]

76. Bodory, H.; Camponovo, L.; Huber, M.; Lechner, M. The finite sample performance of inference methods for propensity score matching and weighting estimators. *J. Bus. Econ. Stat.* **2020**, *38*, 183–200. [CrossRef]

77. Yin, P.; Fan, X. Estimating R 2 shrinkage in multiple regression: A comparison of different analytical methods. *J. Exp. Educ.* **2001**, *69*, 203–224. [CrossRef]

78. Sharir, G.; Noy, A.; Zelnik-Manor, L. An image is worth 16×16 words, what is a video worth? *arXiv* **2021**, arXiv:2103.13915.