Revue Internationale
de Géomatique

Tech Science Press

DOI: 10.32604/rig.2024.055752

ARTICLE

Using Machine Learning to Determine the Efficacy of Socio-Economic Indicators as Predictors for Flood Risk in London

Grace Gau[1] **and Minerva Singh**[2,3,*]

[1]Department of Earth Science and Engineering, Imperial College London, London, SW7 1NE, UK

[2]Centre for Environmental Policy, Imperial College London, London, SW7 1NE, UK

[3]Nature Based Solutions Initiative (NBSI), School of Geography and Environment, Oxford University, Oxford, SW7 2UA, UK

*Corresponding Author: Minerva Singh. Email: minerva.singh07@imperial.ac.uk

Received: 06 July 2024 Accepted: 12 September 2024 Published: 11 October 2024

ABSTRACT

This study examines how socio-economic characteristics predict flood risk in London, England, using machine learning algorithms. The socio-economic variables considered included race, employment, crime and poverty measures. A stacked generalization (SG) model combines random forest (RF), support vector machine (SVM), and XGBoost. Binary classification issues employ RF as the basis model and SVM as the meta-model. In multiclass classification problems, RF and SVM are base models while XGBoost is meta-model. The study utilizes flood risk labels for London areas and census data to train these models. This study found that SVM performs well in binary classifications with an accuracy rate of 0.60 and an area under the curve of 0.62. XGBoost outperforms other multiclass classification methods with 0.62 accuracy. Multiclass algorithms may perform similarly to binary classification jobs due to reduced data complexity when combining classes. The statistical significance of the result underscores their robustness, respectively. The findings reveal a significant correlation between flood risk and socio-economic factors, emphasizing the importance of these variables in predicting flood susceptibility. These results have important implications for disaster relief management and future research should focus on refining these models to improve predictive accuracy and exploring socio-economic factors.

KEYWORDS

Machine learning; socioeconomic indicators; flood risk assessment; London; predictive modelling

1 Introduction

Rivers have always served as focal points for human settlements because of the rich resources they provide. Nevertheless, being near rivers also entails the potential hazard of floods. Floods are the most widespread and lethal natural catastrophe, resulting in more than 2 billion fatalities from 1998 to 2017 [1]. According to current projection models, there is an expected rise in the occurrence and intensity of flood events due to climate change [1,2]. London, renowned for its ancient origins dating back to 100 CE, has profited from its strategic position along the riverfront [3]. Nevertheless, it has also been adversely affected by several expensive floods, most notably the catastrophic Great North Sea flood of 1953, resulting in the loss of more than 300 people. The construction of the Thames Barrier

Copyright © 2024 The Authors. Published by Tech Science Press.
This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

and other flood mitigation measures was a direct reaction to this disaster, providing protection for about 500,000 houses situated along the river [4,5]. Flood defenses offer the greatest level of protection for areas near rivers, which typically have higher property values. However, research has shown that predominantly Black communities will be disproportionately affected by flooding. Additionally, the number of affordable housing units at risk of flooding will triple by 2050 [6,7].

Low-income populations have a greater challenge in recovering from economic flood consequences, such as property destruction, owing to their limited disposable income [8]. This is further compounded by their higher susceptibility to flood risk. In England, there is a comparable situation where socially impoverished populations are more vulnerable to floods, although this issue has not been well investigated [9]. Understanding and addressing the socioeconomic disparity is vital to effectively mitigate the probability of flooding in various areas. London is an excellent location for studying due to its varied socioeconomic environment.

Although there is a correlation between socioeconomic class and flood risk, most research studies that investigate flood risk using machine learning algorithms mostly use climatic factors. By using conventional flood indicators like rainfall, elevation, topography, among others, machine learning models may achieve Area Under the Curve (AUC) values above 0.9 in certain regions [10–13]. By integrating geomorphic and socio-economic variables, it is possible to get very precise outcomes with an AUC of 0.88 [14]. Within flood risk studies, the Support Vector Machine (SVM) has been extensively examined, whereas the Random Forest (RF) regularly proves to be the most efficient model, producing the most favorable outcomes in many research publications [15–19].

Prior research on socioeconomic disparities has primarily concentrated on the economic consequences of flooding. This research has taken a reactive approach by examining how populations react to and recover from flood events. However, there has been limited exploration of the predictive capabilities of socioeconomic indicators, which could be crucial in implementing proactive flood mitigation strategies in vulnerable communities [20]. In addition, while there is a correlation between socioeconomic characteristics and flooding, the majority of machine learning research primarily concentrates on environmental predictors. The study on socio-economic aspects has not harnessed the potential of machine learning, instead relying on weather and economic models to draw their results [21,22].

A study conducted by Chen et al. in 2019 employs a hybrid machine learning approach to assess flood risk in urban areas [13]. The researchers utilize Random Forest (RF) and Support Vector Machine (SVM) algorithms, integrating various environmental factors such as rainfall, land use, and topography, alongside socioeconomic indicators like income levels and population density. Their findings indicate that incorporating socioeconomic data significantly improves the model's predictive accuracy, demonstrating the importance of considering these factors in flood risk assessments. Another study conducted by Deroliya et al. in 2022 focuses on the application of deep learning techniques for flood prediction in river basins [14]. The authors employ a Convolutional Neural Network (CNN) model to analyze satellite imagery and hydrological data, aiming to predict flood events more accurately. Their results highlight the effectiveness of deep learning in capturing complex spatial patterns related to flooding, suggesting that integrating advanced machine learning techniques can enhance flood prediction capabilities.

The study addresses a critical gap in understanding the relationship between socioeconomic factors and flood risk in London, which has been underexplored in existing literature. While previous research has acknowledged the correlation between socioeconomic status and flood vulnerability, it has primarily focused on environmental predictors, such as climatic conditions, neglecting the

potential of socioeconomic indicators as predictors of flood risk. Most studies have taken a reactive approach, analyzing how communities respond to flooding rather than proactively identifying at-risk areas based on socioeconomic data. This lack of proactive research limits the ability to implement effective flood mitigation strategies. Furthermore, while there is evidence that marginalized communities are disproportionately affected by flooding, the integration of socioeconomic factors into predictive models remains limited.

The evidence presented above indicates a potentially substantial correlation between socioeconomic indicators and flood risk, leading to the main inquiry of this study: Can an area's socioeconomic status be used as a dependable predictor for flood risk? The study aims to address the gaps between socio-economic indicators and flood risk by investigating whether an area's socio-economics status can reliably predict flood risk. We construct machine learning models that incorporate socio-economics factor to identify the critical variables for predicting flood susceptibility. By doing so, this study expects to provide valuable insight for disaster relief management and contribute to a more equitable approach to flood risk mitigation.

2 Materials and Methods

2.1 Study Area

London, England, is a vast city covering an area of 607 square miles and housing over 8.8 million inhabitants. The city is partitioned into 32 boroughs, with 17 of them adjacent to the River Thames, which serves as the principal waterway of the metropolis. The flood likelihood by rivers and sea in London is depicted in Fig. 1. Additionally, there are 24 boroughs that have smaller rivers flowing into the main river, and 15 of these boroughs have canals inside their bounds. This interconnected system of waterways is known as the Blue-Ribbon Network [23].

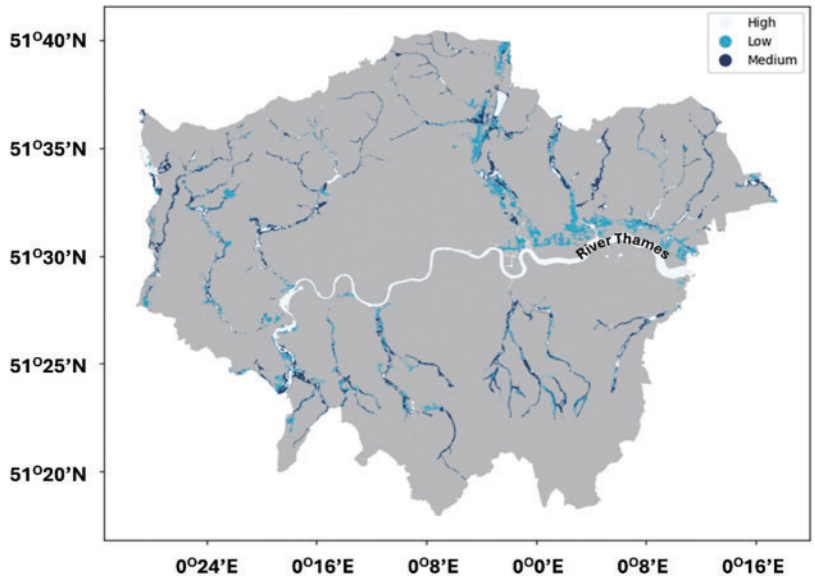

Figure 1: A map of London showing areas of 'High,' 'Medium,' and 'Low' risk for flooding. The grey area represents the entire city of London. Since the flood risk dataset only contains flood risk by rivers and sea, there are no flood risk measures in sites far from waterways

There are several ways to define geographic borders in the city, with the biggest being divided into boroughs, then wards, and lastly, the smallest being a London super output area (LSOA). The average population inside each LSOA border is 1722 individuals, which is the specific data scale used in this study.

2.2 Data

The flood risk data is obtained from the Department for Environment, Food & Rural Affairs of the United Kingdom. The dataset contains a total of 17,223 polygons. The average area of these polygons is 13,054 square meters before anomalies are removed, and 6217 square meters after removal.

The Ordnance Survey's code-point data includes codes for several area borders in London, including LSOA, borough, and ward codes. This dataset includes geometric point data that corresponds to each postcode. The dataset contains a total of 33 boroughs, 654 wards, and 5835 LSOAs. The LSOA atlas available on the London Datastore includes data from the 2011 census for a total of 4766 distinct LSOA codes.

Due to the diverse array of techniques available for gathering geographical data, the process of constructing a unified operational data frame becomes complex. This research thoroughly examines several data modification approaches to mitigate the risk of data loss. Despite the availability of the 'sjoin' function in the geopandas library, which allows for the merging of geometric data based on the inclusion of a point inside a polygon, difficulties arise when dealing with intersecting polygons and considerable variations in polygon size. Every polygon is surrounded by another polygon, which may have a different label. The blue polygon is classified as having a 'High' risk label, whereas the green polygon is classified as having a 'Very Low' label. If the 'sjoin' function were to be used, there would be a significant degree of uncertainty since points that fall on boundary lines might potentially be assigned to the label of the neighboring polygon. This dataset has more than 7,000,000 crossing points, which further exacerbates this problem.

Furthermore, the significant difference in polygon widths and the unequal concentration of postcodes result in an unbalanced distribution, where some labels are either overrepresented or underrepresented. Postcodes are allocated according to the density of addresses in a certain location, resulting in a greater number of post-codes near the city core. To address these problems, the custom function 'nearest' is used to combine geographical data by considering their closeness.

This research eliminates the classification of 'Very Low' for three main reasons. The 'Very Low' classification is defined by an abnormally large average area that surpasses 82,000 square meters. These examples are outliers, diverging from the patterns identified in other labels, each with means below 8000 square meters. Furthermore, the regions classified as 'Very Low' risk are mostly located around the River Thames, which is significantly influenced by flood protection measures such as the Thames Barrier. Zone 1 flood regions are defined as locations having a probability of flooding less than 1 in 1000 per year. Zone 1 sites do not need additional clearances or site specific flood risk studies for home construction locations, unlike Zones 2 and 3 [24]. Considering these criteria, eliminating this label improves model accuracy by ensuring that the centroid points produced are more representative of the polygons contained in the dataset, since big polygons are excluded. In addition, eliminating this label enhances the model's accuracy in reflecting state decisions that designate some regions of flood risk as suitable for development.

The selection of socioeconomic status indicators is made with careful deliberation. Research conducted in the United States has shown that communities of color and low income areas are disproportionately affected by flood risk [6]. Therefore, in this work, socioeconomic parameters

studied include racial distribution percentages and other income status indicators such as mean home price, free school lunch rates, job rates, and income assistance rates. Racial distribution percentages capture the proportion of different racial groups within each area, highlighting that communities of color often face higher flood risks due to historical and socio-economic disparities. Mean home price serves as a proxy for wealth and economic stability, with higher-priced areas typically having better infrastructure and resources to mitigate flood risks. Free school lunch rates, a common indicator of child poverty, help identify low-income areas (see Fig. 2a) where families might lack the financial resources to recover from flood damage. Job rates indicate the economic health of an area, with higher employment levels suggesting better economic conditions that can influence the ability to implement and maintain flood defenses. Income assistance rates reflect the proportion of the population receiving government financial aid, identifying economically disadvantaged areas that may struggle with flood recovery. As shown in Fig. 2b, the mean house price in the top righthand corner that higher house prices are concentrated in central London where there is no flood risk in Fig. 1, while Crimes and BAME (Fig. 2c,d) shows that there is a stronger concentration of POC living outside of this centralized area. The merging of these indicators is accomplished by using LSOA boundary codes in conjunction with postcode and flood risk data, resulting in the creation of a complete dataset (Fig. 3). LSOAs located far away from waterways are not included in the flood risk statistics since they only cover flood danger from rivers and the sea. Hence, the final dataset has 1150 distinct LSOA codes.

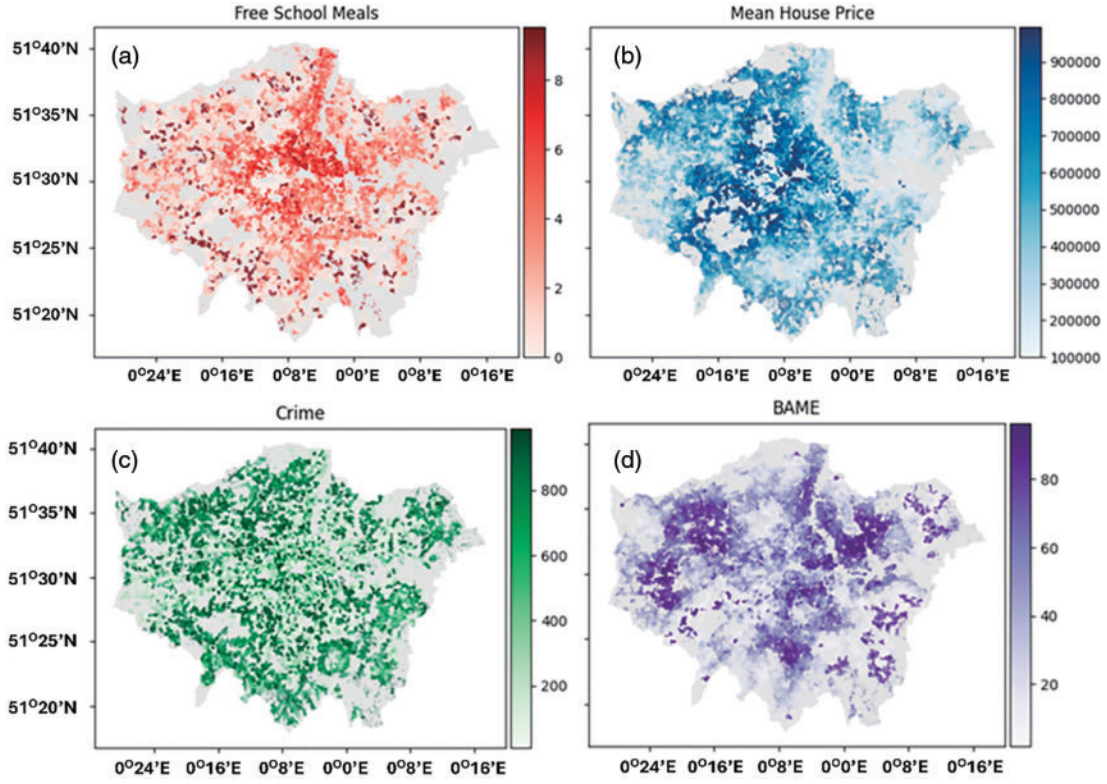

Figure 2: Heatmap of the distribution of different socioeconomic indicators throughout London. Percentage of (a) free school meals, (b) mean house prices, (c) notifiable offences, and (d) Black, Asian, and Minority Ethnic (BAME) per LSOA

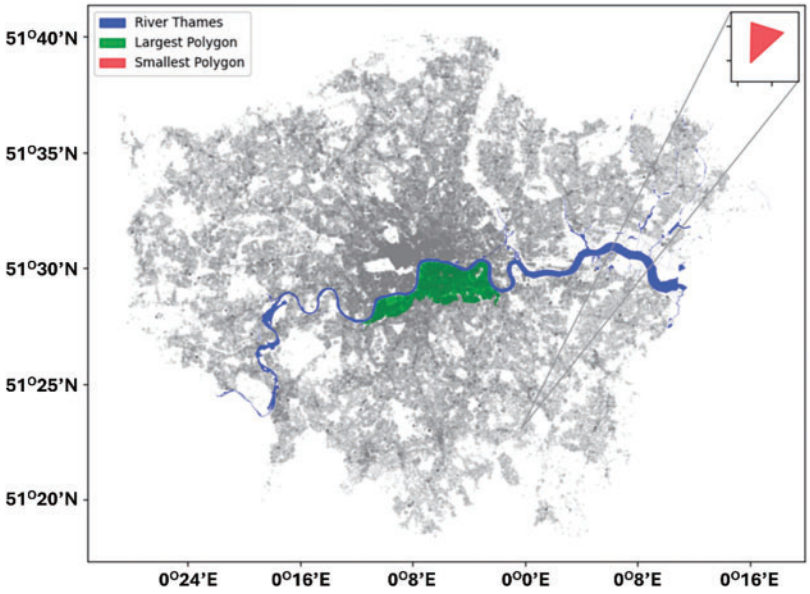

Figure 3: Map showing the geographical boundaries of all postcodes in London in the color gray. Additionally, include the biggest and smallest polygons from the flood risk information. The River Thames is categorized as having a 'High' danger of flooding, whereas the green polygon is categorized as having a 'Very Low' risk of flooding. The zoom box that encompasses the tiniest polygon is enlarged to a scale of 120,000 times its original size

2.3 Algorithms

Two major challenges in this work are model overfitting and the complexities of label imbalance. To surmount these obstacles, a variety of methods are used. The high level of detail in this data significantly leads to overfitting. Each data point in the dataset is assigned a unique LSOA code. Therefore, if the data is randomly divided into training and validation sets, it is possible for neighboring points to be separated into different groups. This may lead to the model memorizing properties of closely related points. To handle this, distinct ward codes are chosen at random and used to partition the data into training and validation sets, guaranteeing that no data with the same ward code is distributed throughout the datasets.

Grid Search is an effective method for tuning hyperparameters, since it methodically explores various combinations of hyperparameters to identify the values that provide the most precise outcomes. Every model has a distinct collection of hyperparameters that may be adjusted to enhance accuracy and mitigate overfitting. Paying careful attention to detail is necessary while conducting grid search owing to its heightened sensitivity to certain characteristics. As an example, with the XGBoost algorithm, increasing the learning rate from 0.3 to 0.4 resulted in a six percent improvement in training accuracy.

Sequential Forward Selection (SFS) is a strategy that effectively decreases spatial overfitting [25]. The SFS algorithm systematically incorporates individual characteristics into an initially empty collection of features to evaluate their influence on the performance of the model. This technique eliminates extraneous input and interference, enabling the model to be only trained on data that enhances outcomes.

SMOTE, or Synthetic Minority Oversampling Technique, is used to address the problem of unbalanced data by generating synthetic samples of the underrepresented class. SMOTE can produce fresh samples using several methods. RF effectively catches intricate data, which is why Borderline SMOTE is used. Borderline SMOTE is a technique that resamples data points located at the border of several labels. This allows models to capture more detailed data, since these points are more likely to be misclassified [26] SVM SMOTE is used in SVM to produce synthetic data near the borders, hence enhancing the representation of the optimum decision boundary in the final model.

The SVM is a supervised learning technique mostly used for classification tasks. It has impressive performance on small datasets because to its ability to categorize data points and identify a hyperplane, or decision boundary, that efficiently separates the data into different classes while maximizing the distance between these classes [27,28]. SVM inherently mitigates overfitting by seeking the most resilient hyperplane, hence avoiding the model from gathering extraneous data.

RF is a popular ensemble learning method that constructs numerous Decision Trees using different subsets of the data and characteristics. The predictions of these trees are then combined to provide the final predictions. The forecasts are determined using a voting process, where the ultimate prediction is the class that obtains the greatest number of votes [29]. This approach enhances the precision of decision tree models by mitigating the issue of overfitting, which decision trees are prone to [30]. This report [31] is recognized for its capacity to decrease high-dimensional, multisource data.

XGBoost, also known as Extreme Gradient Boosting, is an ensemble learning approach, like RF. Sequentially building numerous weak decision trees is very advantageous when dealing with huge and complicated datasets. Each succeeding tree in the sequence corrects mistakes made by the previous model, allowing for the capture of intricate details in the data [32].

SFS is ineffective in XGBoost because of the algorithms built-in ‘greedy’ optimization, which chooses the best feature for each tree and the built-in L2 regularization, which penalizes models that are too complex [33]. Therefore, when SFS was applied to the model, results were not improved, and the decision to exclude SFS in the final model was made.

Stacked Generalization is a kind of ensemble learning approach that merges many models together to enhance performance. Stacking is a technique that involves using the outputs of base models, which have been trained on the training data, and feeding them into the meta-model [34]. This research chooses base and meta-models depending on the performance of each unique model.

When it comes to binary classification, the RF model is the most suitable option since it excels at capturing intricate associations. The precision and capacity of SVM to address overfitting make it a commendable meta-model. RF and SVM are used as basis models in multiclass settings, whereas XGBoost is used as the meta-model because of its effectiveness in dealing with unbalanced data.

The chosen methodologies—SVM, RF, XGBoost, and Stacked Generalization—were selected for their strengths in handling complex, high-dimensional data and their ability to improve predictive accuracy through ensemble learning. These models can significantly impact real-world flood prediction by providing more accurate and robust assessments of flood risk, especially when integrating socio-economic factors. This can help disaster relief officials and urban planners allocate resources more effectively, ensuring that high-risk areas, particularly those with vulnerable populations, receive the necessary support and interventions. Additionally, these models can inform policies aimed at reducing social disparities in disaster preparedness and response, promoting a more equitable approach to flood risk management.

Models are assessed using many performance measures, including precision, recall, F1 scores, accuracy in training, validation, and cross validation, and AUC score in binary classification tasks. The evaluation of precision, recall, and F1 is conducted on the validation dataset. The scores for all eight trained models may be seen in Table 1.

Table 1: Validation metrics for each model with the best score for each bolder. Train accuracy rankings are determined by how similar the train accuracy is to validation and cross validation scores with $p < 0.05$ for each model

Multiclass classification (High, Medium, and Low risk)				
	SVM	RF	XGBoost	Stack Gen.
Precision	0.60	0.61	0.62	0.61
Recall	0.58	0.58	0.59	0.59
F1	0.52	0.53	0.54	0.51
Train accuracy	0.59	0.79	0.61	0.63
Validation accuracy	0.58	0.58	0.59	0.59
Cross validation	0.55	0.58	0.57	0.55
Binary classification (High, Medium, and Low risk)				
Precision	0.60	0.58	0.57	0.58
Recall	0.60	0.57	0.56	0.57
F1	0.60	0.57	0.56	0.57
Train accuracy	0.67	0.89	0.93	0.85
Validation accuracy	0.60	0.57	0.56	0.57
Cross validation	0.60	0.61	0.57	0.65
AUC	0.62	0.61	0.57	0.59

Precision is a metric that highlights the occurrence of erroneous positives, while recall concentrates on false negatives. F1 is a metric that seeks to achieve a balance between both [35,36]. Recall is prioritized because it measures the proportion of actual high-risk flood areas correctly identified, which is crucial to minimize false negatives. False negatives, where high-risk areas are mislabeled as low risk, can have severe consequences, such as leaving properties uninsured or unprotected. While precision is important for reducing false positives, recall's role in accurately identifying high-risk areas is more critical in this context. The F1 score balances precision and recall, but recall remains the primary focus due to the severe impact of false negatives. Although accuracy provides an overall correctness measure, it can be misleading in class-imbalanced datasets typical in flood risk prediction. AUC summarizes the model's discrimination ability but does not specifically balance precision and recall. By prioritizing recall, the models robustly predict vulnerable areas, enhancing flood risk management strategies.

$$\text{Precision} = (\text{True Positives})/(\text{True Positives} + \text{False Positives}) \tag{1}$$

$$\text{Recall} = (\text{True Positives})/(\text{True Positives} + \text{False Negatives}) \tag{2}$$

$$\text{F1} = 2 * (\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall}) \tag{3}$$

Additionally, the data were analyzed using Student's *t*-test for two-group comparisons, or a one-way ANOVA for multiple-group comparisons. The statistical significance was set at $p < 0.05$.

3 Results

3.1 Multiclass Classification

XGBoost surpasses all other models in five out of six validation criteria, attaining precision, recall, and validation accuracy scores of 0.6 or above. This model successfully mitigates the problem of overfitting on the training data and demonstrates good generalization to new, unknown data. This is supported by the validation accuracy score of 0.6 and the cross validation score of 0.57. Although SVM has impressive accuracy and recall scores of 0.60 and 0.58, respectively, and effectively avoids overfitting on the training data, it does not do as well in generalizing to new, unknown data compared to other models.

Although the training data is prone to overfitting, RF still benefits from the use of grid search and Sequential Forward Selection (SFS) implementation. These techniques help to decrease the training accuracy from 1 to 0.78. In addition, these strategies enhance memory from 0.44 to 0.58, demonstrating their effectiveness.

3.2 Binary Classification

In the context of binary classification, SVM provides superior performance compared to other methods in six out of seven measures. Although all models had similar precision and recall scores, SVM stands out for its superior performance in reducing train overfitting, achieving a final train accuracy of 0.67. In addition, the SVM has a maximum AUC value of 0.62. The cross validation score of the SG model indicates its impressive performance on new and unseen data, demonstrating its resilience.

In general, binary models exhibit similar levels of accuracy and recall as multiclass models, but they achieve higher F1 scores, because of the balanced distribution of classes in the binary dataset.

3.3 Feature Importance

The SFS findings (Table 2) emphasize the importance of racial distribution as a crucial characteristic, with BAME, white, mixed, and other ethnicities (excluding Black, White, Asian, or Mixed) being used in all algorithms. The mean property price is consistently chosen as a crucial factor in all models. Only two out of the eight final models exclude total notifiable criminal offences and homes with no adults in work. The second most often used element is free school lunches, which is a widely used indication of low income households. Fig. 4 illustrates the distribution of these characteristics across London.

Table 2: Optimal metrics for each model. Note that 'All Features' means every feature in the key below is used in the model. N/A indicates that the model tuning technique was not implemented

Final models			
Model	Grid search parameters	SFS features	SMOTE
SVM (multi)	C: 1, kernel: 'rbf', gamma: 'scale'	11–15 FSM, Price, BAME, White, Mixed, Asian, Black, Other	SVM

(Continued)

Table 2 (continued)

Final models			
Model	Grid search parameters	SFS features	SMOTE
RF (multi)	max_depth: 8, max_features: 'sqrt', min_samples_leaf: 6, min_samples_split: 2, n_estmators: 60	CP, COW, 5–10 FSM, Crime, NAE, Price, BAME, White, Mixed, Black, Other	Borderline
XGBoost (multi)	gamma: 6, learning_rate: 0.4, max_depth: 5, min_child_weight: 7, n_estimators: 5	N/A	N/A
Stack Gen. (multi)	rf_n_estimators: 75, rf_min_samples_split: 4, rf_min_samples_leaf: 8, rf_max_depth: 6, svm_C: 2, xgb_learning_rate: 0.4, xgb_max_depth: 7, xgb_min_child_weight: 9, xgb_n_estimators: 10	CP, COW, 5–10 FSM, Crime, NAE, Price, BAME, White, Mixed, Black, Other	Borderline
SVM (binary)	C: 1, kernel: 'rbf', degree: 3, gamma: 'scale' (default parameters)	N/A	N/A
RF (binary)	max_depth: 18, max_features: 'sqrt', min_samples_leaf: 6, min_samples_split: 2, n_estmators: 87	NAE, Price, BAME, White, Mixed, Asian, Other	N/A
XGBoost (binary)	gamma: 6, learning_rate: 2, max_depth: 4, min_child_weight: 1, scale_pos_weight: None	N/A	N/A
Stack Gen. (binary)	C: 2, degree: 1, gamma: 'scale', kernel: 'rbf', max_depth: 6, max_features: 'sqrt', min_samples_leaf: 4, min_samples_split: 2, n_estimators: 50	5–10 FSM, 11-15 FSM, Crime, NAE, Price, BAME, White, Asian, Other	N/A

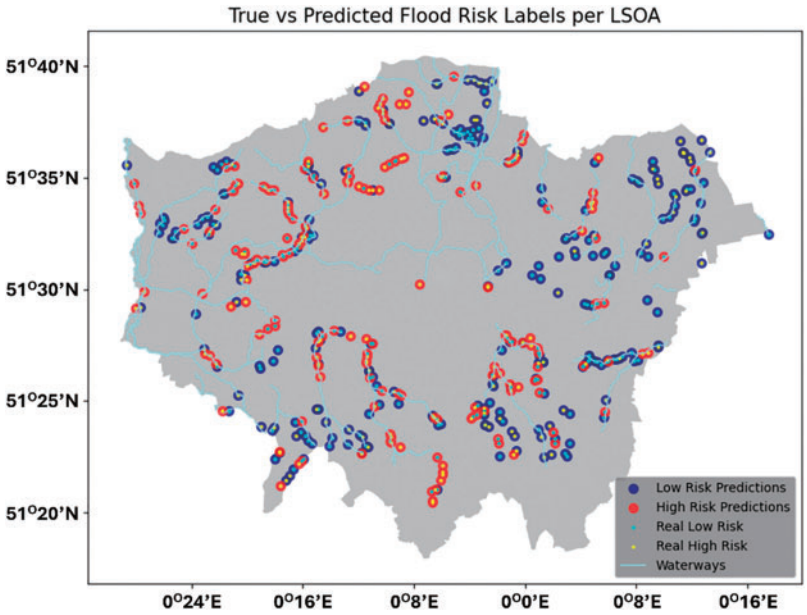

Figure 4: A map of high and low flood risk predictions with true risk labels for each LSOA area in London was created using predictions from the binary SVM model

4 Discussion

4.1 Key Findings

The ML models' capacity to accurately forecast flood risk labels, achieving precision scores of up to 0.62, demonstrates that the socioeconomic condition of a region may serve as a dependable feature for flood risk accuracy, even without the inclusion of environmental factors. Studies that forecast floods based on environmental characteristics exhibit significant variability in their outcomes. Certain models get SVM recall scores of 0.5, while others can achieve higher recall ratings of 0.67 via the use of hybrid prediction methods [37]. The findings of this study demonstrate that the association between socioeconomic level and flood risk is extremely significant, as shown by the SVM recall of 0.6. However, it is important to note that the quality of data in research sites influences these results.

The use of SFS to the models in this research yields a consistent list of variables that enhance model outcomes. The most often chosen parameters were racial distribution, housing costs, crime rates, households with no employed adults, and the percentage of students receiving free school meals. Curiously, the income variables used by SFS in all models are measurements of poverty experienced by families or throughout children. Childhood poverty and children receiving out-of-work benefits are the primary factors that significantly impact the model's performance in RF and SG. SFS does not include statistics such as the rate of individuals dependent on income assistance and unemployment rates, which are measures of adult poverty.

The Random Forest (RF) and Support Vector Machine (SVM) models demonstrated strong performance, particularly in binary classification tasks, with SVM achieving an accuracy rate of 0.60. RF, known for its robustness in handling complex datasets, performed well in both binary and multiclass classifications, effectively capturing intricate relationships between socioeconomic indicators and flood risk. In contrast, the XGBoost model outperformed others in multiclass classification with

an accuracy of 0.62, highlighting its ability to manage varying data complexities and interactions among features. The superior performance of XGBoost can be attributed to its gradient boosting framework, which optimizes model accuracy through iterative learning and can handle missing data more effectively. However, the performance of these models can vary based on the conditions under which they are applied, such as data distribution and the presence of noise. For instance, while SVM is effective in high-dimensional spaces, it may struggle with larger datasets due to increased computational demands. By analyzing these strengths and weaknesses, we can gain valuable insights into the conditions that favor each model, guiding future research and practical applications in flood risk assessment. This detailed comparison will enhance our understanding of model performance and inform the selection of appropriate algorithms for similar studies in different contexts.

Additional investigation is necessary to establish the cause of these findings, but plausible factors may include the need for families to have greater space to raise children. Families are more inclined to choose bigger living areas for the purpose of comfort and the overall welfare of their children. Research has shown that children who are raised in overcrowded households have substantial adverse effects. Nevertheless, bigger homes command higher prices, therefore prompting the search for locations with lower property values [38]. In addition, the financial burden of raising children is substantial, ranging from £157,000 to £208,000 for the whole period from infancy to age 21. This expense has the potential to push families farther into poverty [39]. Another potential factor influencing the outcomes of this research might be the higher prevalence of child poverty compared to adult poverty. London has a child poverty rate of 34% among children aged 5–9. By contrast, a mere 17% of individuals between the ages of 30 and 34 [40]. This study reveals the socioeconomic characteristics that have a direct influence on flood risk prediction. However, more research is necessary to fully comprehend the reasons behind the significance of each element.

4.2 Challenges

The primary obstacle of this work was the creation of a thorough and precise dataset. Due to the rarity of census reports and the presence of discrepancies in geographic data collection, it is necessary to do data modification to construct a functional dataframe. By performing certain operations, such as combining geographic dataframes based on the closest neighboring points, it is not feasible to preserve the whole of the original data. However, methods such using a customized ‘nearest’ function are used to reduce the loss of information.

Geospatial information is susceptible to overfitting due to the dense clustering of data points. It is necessary to apply measures to prevent models from memorizing closely similar points. One way to do this is by dividing the training and validation data based on distinct ward numbers. The unbalanced nature of flood risk data makes it susceptible to overfitting. The ‘Low’ label was assigned to more than 53% of the final dataset, but the ‘High’ label was assigned to just 11%. SMOTE and hyper parameter adjustment effectively addresses the issue of uneven distribution, while some models may still exhibit small overfitting on the training data.

4.3 Limitations of the Research

The use of 2011 census data, because of the lack of 2021 data, is a significant constraint. Furthermore, this research does not take into consideration the recent modification of LSOA borders that were made to accommodate changes in population distribution. The limited emphasis on inundation caused by rivers and the sea in the flood dataset limits the extent of this study since it fails to include regions that are far from watercourses (Fig. 1). Out of the 4766 distinct LSOA codes in the socioeconomic dataset, only 1150 codes overlap with the flood data. The replication of this research

should be undertaken during the winter of 2023, coinciding with the publication of the definitive census results. Moreover, including data on floods caused by groundwater, surface water, and other origins may enhance the resilience and effectiveness of models.

London, with its diverse socioeconomic landscape, presents distinct challenges and characteristics that may not be applicable to other regions. For instance, previous studies have highlighted how urban density, historical infrastructure, and varying levels of flood defense measures can significantly impact flood risk assessments. By comparing our findings with similar studies conducted in other urban settings, we can better understand how different environmental and socioeconomic contexts influence flood risk outcomes. This comparative analysis will not only enhance the robustness of our conclusions but also provide valuable insights into how flood risk management strategies can be tailored to address the specific needs of different communities. Therefore, future discussions should incorporate a broader perspective, considering how the unique attributes of London, as well as findings from other studies, can inform more effective and equitable flood risk mitigation strategies.

4.4 Future Research

Examining the timing of the implementation of construction restrictions and flood laws may provide valuable information about the effectiveness of measures aimed at reducing flood risk. Moreover, by combining historical flood data with past census data, it is possible to determine whether policies have strengthened the correlation between socioeconomic position and the likelihood of experiencing floods.

Although this research specifically examines London, England, it is important to note that income inequality affecting communities' susceptibility to floods is not limited to this region. Extending to further cities in England and nationwide can reveal more complex and distinctive patterns. Analyzing outcomes from various cities may assist in assessing the efficacy of flood regulations that have been applied nationwide. On 31 August 2023, London Datastore revised the dataset titled 'Children in low income families'. Utilizing this information for training machine learning models might strengthen the conclusions obtained in this paper about the significance of childhood poverty in forecasting flood risk labels.

Using just socioeconomic data to evaluate flood risk is insufficient in creating a flawless model, since some places are not prone to floods owing to natural factors such as elevation and geography. Nevertheless, the findings of this research indicate that socioeconomic characteristics have the potential to enhance the performance of flood risk models that are trained using environmental data.

5 Conclusions

This study aims to initiate a discussion in government settings on strategies to reduce the negative effects of floods on vulnerable communities and to bring attention to the existing socioeconomic inequalities. Given the projected rise in both the destructive power and occurrence of flood occurrences due to climate change, it is important to have a comprehensive understanding of the societal consequences of floods. This information may be used to develop more effective defensive measures against such disasters. In 2021, a total of £5.2 billion was allocated to a flood defense program aimed at safeguarding the wellbeing of more than 300,000 individuals [41,42]. This unequivocally demonstrates the government's commitment to safeguarding civilians from the perils of floods. The data collected from this research can guarantee that the significant expenditures being made are most advantageous, by directing resources to the most vulnerable populations.

The models trained in this research have shown the capability to forecast flood labels with an accuracy of up to 0.62, even without using any environmental characteristics. This finding supports the integration of socioeconomic factors into flood risk assessments, encouraging government agencies to adopt policies that address social disparities in disaster preparedness. Specifically, the findings suggest several policy actions: directing flood defense funding and resources to high-risk areas identified by socioeconomic indicators, developing inclusive disaster preparedness programs for low-income and minority communities, revising flood insurance policies to provide better coverage for disadvantaged populations, and integrating flood risk predictions into urban planning to enhance infrastructure in vulnerable areas. By implementing these recommendations, policymakers can create a fairer and more just approach to flood risk management, addressing both immediate and long-term needs of vulnerable populations. This research underscores the importance of proactive and inclusive policies to enhance resilience against the growing threat of climate-induced flooding.

Acknowledgement: The authors are grateful for the support of the Center for Environmental Policy at Imperial College London.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Grace Gau and Minerva Singh; Data curation, Grace Gau; Formal analysis, Grace Gau; Investigation, Grace Gau; Methodology, Minerva Singh and Grace Gau; Project administration, Minerva Singh; Resources, Minerva Singh; Supervision, Minerva Singh; Writing—original draft, Grace Gau; Writing—review and editing, Minerva Singh and Grace Gau. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Minerva Singh, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. World Health Organization (WHO). Floods; 2019. Available from: https://www.who.int/health-topics/floods. [Accessed 2023].
2. Otto FE, van der Wiel K, van Oldenborgh GJ, Philip S, Kew SF, Uhe P, et al. Climate change increases the probability of heavy rains in Northern England/Southern Scotland like those of storm Desmond—a real-time event attribution revisited. Environ Res Lett. 2018 Jan 29;13(2):024006. doi:10.1088/1748-9326/aa9663.
3. National Geographic. Understanding rivers; 2023. Available from: https://education.nationalgeographic.org/resource/understanding-rivers/. [Accessed 2023].
4. Environment Agency. The Thames Barrier. United Kingdom Government; 2023. Available from: https://www.gov.uk/guidance/the-thames-barrier. [Accessed 2023].
5. London City Hall. Flood risks in London; 2014. Available from: https://www.london.gov.uk/about-us/about-us/london-assembly/london-assembly-publications/flood-risks-london3. [Accessed 2023].

6. Wing OE, Lehman W, Bates PD, Sampson CC, Quinn N, Smith AM, et al. Inequitable patterns of US flood risk in the Anthropocene. Nat Clim Change. 2022 Feb;12(2):156–62. doi:10.1038/s41558-021-01265-6.
7. Buchanan MK, Kulp S, Cushing L, Morello-Frosch R, Nedwick T, Strauss B. Sea level rise and coastal flooding threaten affordable housing. Environ Res Lett. 2020 Dec 1;15(12):124020. doi:10.1088/1748-9326/abb266.
8. PreventionWeb. SENDAI framework for disaster risk reduction 2015–2030. Poverty and inequality; 2021. Available from: https://www.preventionweb.net/understanding-disaster-risk/risk-drivers/poverty-inequality. [Accessed 2023].
9. The City Data Team. London datastore, Greater London Authority. Economic fairness–Wealth inequality; 2020. Available from: https://data.london.gov.uk/economic-fairness/equal-opportunities/wealth-inequality. [Accessed 2023 Oct 1].
10. Habibi A, Delavar MR, Sadeghian MS, Nazari B. Flood susceptibility mapping and assessment using regularized random forest and naïve bayes algorithms. ISPRS Ann Photogramm Remote Sens Spat Inform Sci. 2023 Jan 13;10:241–8. doi:10.5194/isprs-annals-x-4-w1-2022-241-2023.
11. Li Y, Hong H. Modelling flood susceptibility based on deep learning coupling with ensemble learning models. J Environ Manag. 2023 Jan 1;325(21):116450. doi:10.1016/j.jenvman.2022.116450.
12. Knighton J, Buchanan B, Guzman C, Elliott R, White E, Rahm B. Predicting flood insurance claims with hydrologic and socioeconomic demographics via machine learning: exploring the roles of topography, minority populations, and political dissimilarity. J Environ Manag. 2020 Oct 15;272:111051. doi:10.1016/j.jenvman.2020.111051.
13. Chen W, Hong H, Li S, Shahabi H, Wang Y, Wang X, et al. Flood susceptibility modelling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles. J Hydrol. 2019 Aug 1;575:864–73. doi:10.1016/j.jhydrol.2019.05.089.
14. Deroliya P, Ghosh M, Mohanty MP, Ghosh S, Rao KD, Karmakar S. A novel flood risk mapping approach with machine learning considering geomorphic and socio-economic vulnerability dimensions. Sci Total Environ. 2022 Dec 10;851:158002. doi:10.1016/j.scitotenv.2022.158002.
15. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019 Dec;19(1):1–6. doi:10.1186/s12911-019-1004-8.
16. Park SJ, Lee DK. Prediction of coastal flooding risk under climate change impacts in South Korea using machine learning algorithms. Environ Res Lett. 2020 Aug 25;15(9):094052. doi:10.1088/1748-9326/aba5b3.
17. Ennouali Z, Fannassi Y, Lahssini G, Benmohammadi A, Masria A. Mapping Coastal vulnerability using machine learning algorithms: a case study at North coastline of Sebou estuary, Morocco. Reg Stud Mar Sci. 2023 Jun 1;60:102829. doi:10.1016/j.rsma.2023.102829.
18. Hasan MH, Ahmed A, Nafee KM, Hossen MA. Use of machine learning algorithms to assess flood susceptibility in the coastal area of Bangladesh. Ocean Coast Manage. 2023 Apr 1;236:106503. doi:10.1016/j.ocecoaman.2023.106503.
19. Zhang S, Zhang J, Li X, Du X, Zhao T, Hou Q, et al. Estimating the grade of storm surge disaster loss in coastal areas of China via machine learning algorithms. Ecol Indic. 2022 Mar 1;136:108533. doi:10.1016/j.ecolind.2022.108533.
20. Samsuddin A, Kaman ZK, Husin NM. Socio-economic assessment on flood risk impact: a methodological review toward environmental sustainability. IOP Conf Ser: Earth Environ Sci. 2021 Dec 1;943:012010. doi:10.1088/1755-1315/943/1/012010.
21. Tripathy SS, Vittal H, Karmakar S, Ghosh S. Flood risk forecasting at weather to medium range incorporating weather model, topography, socio-economic information and land use exposure. Adv Water Resour. 2020 Dec 1;146:103785. doi:10.1016/j.advwatres.2020.103785.
22. Zeng Z, Guan D, Steenge AE, Xia Y, Mendoza-Tinoco D. Flood footprint assessment: a new approach for flood-induced indirect economic impact measurement and post-flood recovery. J Hydrol. 2019 Dec 1;579:124204. doi:10.1016/j.jhydrol.2019.124204.

23. London City Hall. London waterways; 2016. Available from: https://www.london.gov.uk/programmes-strategies/planning/who-we-work/planning-working-groups/london-waterways. [Accessed 2023].
24. Department for Levelling Up, Housing and Communities. National planning policy framework; 2023. Available from: https://www.gov.uk/government/publications/national-planning-policy-framework–2. [Accessed 2023].
25. Yu H, Wu Y, Niu L, Chai Y, Feng Q, Wang W, et al. A method to avoid spatial overfitting in estimation of grassland above-ground biomass on the Tibetan Plateau. Ecol Indic. 2021 Jun 1;125(2):107450. doi:10.1016/j.ecolind.2021.107450.
26. Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International Conference on Intelligent Computing, 2005 Aug 23; Berlin, Heidelberg: Springer Berlin Heidelberg; p. 878–87. doi:10.1007/11538059_91.
27. Shmilovici A. Support vector machines. In: Data mining and knowledge discovery handbook. Boston, MA, USA: Springer; 2010. p. 231–47.
28. Chand N, Mishra P, Krishna CR, Pilli ES, Govil MC. A comparative analysis of SVM and its stacking with other classification algorithm for intrusion detection. In: 2016 International Conference on Advances in Computing, Communication, & Automation (ICACCA) (Spring), 2016 Sep 8; Dehradun, India: IEEE; p. 1–6.
29. Yang Y. Temporal data mining via unsupervised ensemble learning. 1st ed. Cambridge, MA, USA: Elsevier; 2016 Nov 15.
30. Breiman L. Random forests. Mach Learn. 2001 Oct;45(1):5–32. doi:10.1023/A:1010933404324.
31. Sarica A, Cerasa A, Quattrone A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Front Aging Neurosci. 2017 Oct 6;9:329. doi:10.3389/fnagi.2017.00329.
32. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016 Aug 13; San Francisco, CA, USA; p. 785–94.
33. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. Eur J Epidemiol. 2018 May;33(5):459–64. doi:10.1007/s10654-018-0390-z.
34. Sreeram ASK. Stacking classifier approach for a multi-classification problem; 2021. Available from: https://towardsdatascience.com/stacking-classifier-approach-for-a-multi-classification-problem-56f3d5e120c8. [Accessed 2023].
35. Yu H, Luo Z, Wang L, Ding X, Wang S. Improving the accuracy of flood susceptibility prediction by combining machine learning models and the expanded flood inventory data. Remote Sens. 2023 Jul 19;15(14):3601. doi:10.3390/rs15143601.
36. Ekmekcioğlu Ö., Koc K, Özger M, Işık Z. Exploring the additional value of class imbalance distributions on interpretable flash flood susceptibility prediction in the Black Warrior River basin, Alabama, United States. J Hydrol. 2022 Jul 1;610(2):127877. doi:10.1016/j.jhydrol.2022.127877.
37. Lawal ZK, Yassin H, Zakari RY. Flood prediction using machine learning models: a case study of Kebbi state Nigeria. In: 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2021 Dec 8–10; Brisbane, Australia: IEEE; p. 1–6.
38. Moneyfarm. How much does it cost to raise a child in the UK? 2023. Available from: https://blog.moneyfarm.com/en/financial-planning/how-much-does-it-cost-to-raise-a-child/. [Accessed 2023].
39. Trust for London. Poverty before and after housing costs by age; 2022. Available from: https://trustforlondon.org.uk/data/poverty-by-age-group. [Accessed 2023].

40. Parliament UK. Flood plains: housing development; 2021. Available from: https://hansard.parliament.uk/Lords/2021-06-24/debates/B360AE77-0CE1-4D64-B79D-F878C7E2DB70/FloodPlainsHousingDevelopment. [Accessed 2023].
41. Solari CD, Mare RD. Housing crowding effects on children's wellbeing. Soc Sci Res. 2012 Mar 1;41(2):464–76. doi:10.1016/j.ssresearch.2011.09.012.
42. GeoSmart Information. Flood zones explained; 2023. Available from: https://geosmartinfo.co.uk/2016/03/flood-zones-explained/. [Accessed 2023].