Data-Driven Forecasting of Refugee Displacement Using Machine and Deep Learning Models with XAI

Ibrahim Sani*, Umar Hasan*, Raihan Sharif, Tahfeem Islam Siam, Md Alamgir Hossain, and Riasat Khan
Electrical and Computer Engineering, North South University, Dhaka, Bangladesh
{ibrahim.sani, umar.hasan, raihan.sharif02, tahfeem.siam, hossain.md, riasat.khan}@northsouth.edu
*Equal Contribution

***Abstract*—The global displacement of people due to conflict, persecution, and political instability has reached unprecedented levels, creating significant challenges for humanitarian organizations and host nations. Accurate prediction of refugee and asylum seeker flows is crucial for effective resource allocation and policy-making. This work presents a data-driven approach to predicting the scale of displacement events using a comprehensive United Nations High Commissioner for Refugees (UNHCR) dataset. We extensively evaluate a diverse range of machine learning models, from traditional ensembles to deep learning architectures, including MLP Regressor, TabNet, FT-Transformer, and DistillBERT. A two-level Stacking Regressor combines XGBoost, LightGBM, and AdaBoost as base learners with a Ridge Regressor meta-learner trained on their out-of-fold predictions to enhance final predictive performance. The proposed Stacking Regressor model achieved a remarkable R-squared value of 0.989 on a single holdout test set. A more rigorous 5-fold cross-validation was conducted to assess model generalizability and robustness. These results identified LightGBM as the superior model, achieving the highest average R^2 score of 0.88, demonstrating its reliability and stability against data variance. As a key methodological contribution, we employ a comprehensive Explainable AI (XAI) framework using both LIME and SHAP to interpret model predictions, bridging the critical gap between predictive accuracy and the transparency required for real-world adoption. This work demonstrates the potential of interpretable and robust machine learning to support proactive and evidence-based humanitarian action.**

***Index Terms*—Refugee flow prediction, Machine learning, Deep Learning, Gradient boosting, Explainable AI, Humanitarian action, UNHCR data.**

I. Introduction

Global forced displacement is one of our time's most pressing humanitarian crises. Millions of individuals are compelled to leave their homes due to armed conflict, persecution, political instability, and environmental disasters, seeking safety and asylum in other countries. These large-scale, often unpredictable, migration patterns place immense strain on the resources of host nations and challenge regional stability. Overwhelmed asylum systems can lead to significant processing delays, social tensions, and risks to vulnerable populations [1]. Consequently, accurately forecasting the scale of refugee and asylum seeker flows is paramount for enabling proactive humanitarian aid, informed policy decisions, and efficient resource allocation.

Existing studies have increasingly employed statistical and machine learning (ML) techniques to analyze forced migration data, yielding valuable insights into predicting displacement events and identifying contributing factors [1]. Researchers have demonstrated the effectiveness of ensemble models, i.e., Random Forest and gradient boosting systems, for forecasting, often augmenting official data with non-conventional sources, such as news reports or social media data to capture real-time signals [2]. However, much of the existing work prioritizes predictive accuracy, often at the expense of model transparency. The complex, "black-box" nature of high-performing models can be a significant barrier to their adoption in high-stakes humanitarian contexts, where understanding the rationale behind a prediction is crucial for trust and responsible decision-making [3]. This creates a significant research gap: the need for highly accurate, transparent, and interpretable predictive models in this domain.

To address this research gap, this study makes a key methodological contribution by integrating a comprehensive Explainable AI (XAI) framework with a high-performance predictive pipeline. We present a comparative analysis of multiple machine learning models for predicting the scale of global refugee flows, but argue that accuracy alone is insufficient. By using both Local Interpretable Model-agnostic Explanations (LIME) [4] and SHapley Additive exPlanations (SHAP) [5], we provide transparent, instance-level, and global insights into our best model's predictions, thereby bridging the gap between predictive power and interpretability. The key contributions of this study are:

- A comparative evaluation of a diverse range of machine learning models, from traditional ensembles to deep learning architectures, including MLP Regressor, TabNet, FT-Transformer, and DistillBERT, with hyperparameter optimization for predicting the scale of forced displacement events using a UNHCR dataset.
- Developing a two-level Stacking Regressor, leveraging XGBoost, LightGBM, and AdaBoost as base learners and a Ridge Regressor as the meta-learner, to enhance predictive accuracy by combining their out-of-fold predictions.
- A robust cross-validation of the top-performing models, highlighting the state-of-the-art performance achieved

with a tuned gradient boosting ensemble technique.

- A methodological framework that pairs high-performance predictive modeling with robust XAI techniques (LIME and SHAP) to ensure transparency and trust.
- An analysis of the practical and ethical implications of using predictive models in the humanitarian sector.
- The development of a comprehensive, end-to-end framework for predicting refugee flows that achieves high predictive accuracy while ensuring model transparency and considering the ethical and practical challenges of real-world deployment.

Section II reviews prior research on migration forecasting and explainable AI frameworks. Section III details the employed UNHCR dataset, preprocessing steps, and the machine learning models employed. Section IV presents the experimental setup, performance results, and a discussion of the findings. Section V concludes the paper and suggests directions for future work.

II. Literature Review

A. Machine Learning for Migration Forecasting

The application of ML to forecast migratory flows has gained significant traction. Traditional econometric methods, such as gravity models, often fail to capture the complex, nonlinear dynamics of forced displacement. Inspired by successes in other domains, researchers have adopted more advanced techniques. The work of Boss et al. [1] was seminal, using high-dimensional data and finding that an ensemble of XGBoost and Random Forest models outperformed conventional approaches. Similarly, recent studies have focused on augmenting administrative data with non-conventional, high-frequency data sources to improve predictive power. Santos et al. [2] and Carammia et al. [6] showed that incorporating Google Trends and GDELT event data could enhance forecasting accuracy. Other novel data sources include social media for spatio-temporal analysis and high-resolution satellite imagery for mapping settlement populations [7]. While these studies establish the predictive power of ML, they often leave the model's decision-making process opaque, creating a critical research gap in model interpretability.

B. Ensemble Methods in Predictive Modeling

Ensemble learning, which combines multiple ML models to produce one optimal predictive model, has become a cornerstone of modern data science. Techniques fall into two main categories: bagging methods, including Random Forest, which reduces variance by averaging predictions from models trained on different data subsets and boosting methods, i.e., AdaBoost, XGBoost, and LightGBM, which build models sequentially to correct the errors of their predecessors [8]. The consistent superior performance of ensembles on tabular data, as demonstrated in numerous domains, inspired our decision to benchmark various techniques. Furthermore, as formalized by Wolpert [9], stacking offers a powerful method to combine diverse models by training a meta-learner on their outputs. Our inclusion of an ensemble Stacking Regressor model in predicting global refugee and asylum seeker flows was motivated by the success of such architectures.

C. Explainability in High-Stakes Domains

The "black-box" problem of complex models is a significant challenge, particularly in sensitive domains, e.g., criminal justice, healthcare, and humanitarian aid [10]. A prediction without a rationale is difficult to trust, act upon, or scrutinize for bias [11]. The field of Explainable AI (XAI) aims to address this. Various techniques, such as LIME [4] and SHAP [5], offer model-agnostic methods to explain individual predictions. Ribeiro et al. [4] introduced LIME to build a local, interpretable model around a prediction to explain its outcome. The growing body of literature on XAI [12] highlights a strong consensus that for AI to be used responsibly, it must be transparent. This context inspired us to make it a core methodology component, ensuring our high-performance models are scrutinized.

Prior research establishes the potential of machine learning, particularly ensemble methods, for forecasting forced displacement [1], [2], [6]. Many recent studies focus on augmenting traditional administrative data with high-frequency, non-conventional sources, i.e., Google Trends or GDELT event data, to improve predictive accuracy [1], [2]. However, the application of deep learning to this problem has been comparatively limited. While some studies have explored foundational architectures like the Multi-Layer Perceptron (MLP), more advanced models designed for tabular data, such as TabNet and FT-Transformer, or language models like DistillBERT, have remained largely unexplored in this domain. Furthermore, as highlighted by frameworks, e.g., Pham and Luengo-Oroz [13], a significant research gap persists regarding the interpretability of these high-performing but often opaque models. This lack of transparency is a significant barrier to adoption in the humanitarian sector, where trust and accountability are paramount [14]. Our work addresses this gap by striving for state-of-the-art predictive accuracy and methodologically integrating a robust XAI framework (LIME and SHAP). Our motivation is to provide a blueprint for developing predictive tools that are accurate, transparent, scrutable, and ultimately more trustworthy for humanitarian decision-makers.

III. Methodology

This study follows a structured ML workflow, as illustrated in Fig. 1. The pipeline begins with data acquisition and EDA, followed by feature engineering, model training and optimization, evaluation, and finally, explainability analysis.

A. Problem Formulation

The predictive task is formulated as a supervised regression problem. Given a feature vector $X = \{x_1, x_2, ..., x_p\}$ for each displacement event, where p is the number of features (e.g., country of origin, country of asylum, population type, year, etc.), the goal is to train a model f that can accurately predict the number of 'Individuals' (y), a continuous target

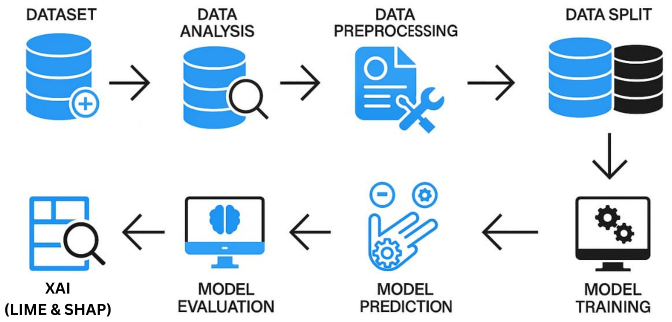


Fig. 1: Workflow for the proposed refugee population displacement prediction system.

variable. The objective is to find a function f that minimizes the prediction error on unseen data:

$$\hat{y} = f(X) \tag{1}$$

where $\hat{y}$ denotes the predicted number of individuals.

B. Data

Dataset: This study utilizes the "UNHCR Situations: Monthly Refugees and Asylum Seekers" dataset from the UNHCR [15]. The raw dataset comprises 506 instances, each detailing a displacement event.

Preprocessing: A rigorous preprocessing pipeline was implemented. Firstly, the 74 missing values in the 'ISO3 of Origin' feature were imputed using the column's mode. Secondly, for the machine learning models, the Interquartile Range (IQR) method was used to identify and remove extreme outliers in the 'Individuals' column to improve model stability. An instance was considered an outlier if its value fell outside the bounds defined by:

$$\text{IQR} = Q_3 - Q_1 \tag{2}$$

$$\text{Upper Bound} = Q_3 + 1.5 \times \text{IQR} \tag{3}$$

$$\text{Lower Bound} = Q_1 - 1.5 \times \text{IQR} \tag{4}$$

For the MLP Regressor, a 'ColumnTransformer' was utilized to apply 'OneHotEncoder' to all categorical features. This created a binary vector representation for each category, a standard practice that prevents the model from inferring a false ordinal relationship. The same transformer also applied 'StandardScaler' to the numerical features.

For the TabNet model, a different approach was taken. Categorical features were converted into numerical representations using 'LabelEncoder'. This method is standard for the TabNet architecture, which is designed to generate its own internal feature embeddings from these integer-based labels during the training process.

C. Exploratory Data Analysis (EDA)

An initial EDA was conducted on the preprocessed data. The dataset is primarily composed of refugees, who constitute 67.8% of the population, with asylum-seekers making up the remaining 32.2%, as shown in Fig. 2. The distribution of the target variable, 'Individuals', is heavily right-skewed, with a mean of 66,706 but a median of only 992, as shown in Fig. 3a. The geographic distribution of events is concentrated, with the Democratic Republic of the Congo being the most frequent country of asylum and Sudan the most frequent country of origin, illustrated in Fig. 4. A time-series plot shows significant activity spikes in 2025, depicted in Fig. 3b. To examine linear relationships, a Pearson correlation matrix was generated (Fig. 5). As expected, the matrix reveals strong positive correlations between related features, i.e., 'Country' and 'ISO3' (0.94). Notably, most features exhibit weak linear correlation with the target variable, 'Individuals,' underscoring the necessity of employing non-linear models.

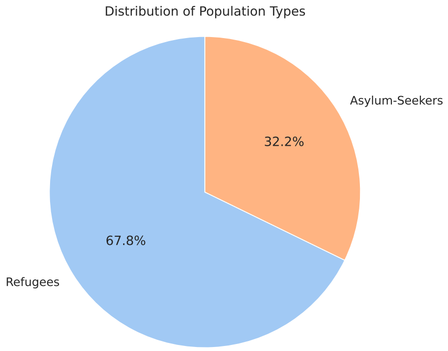


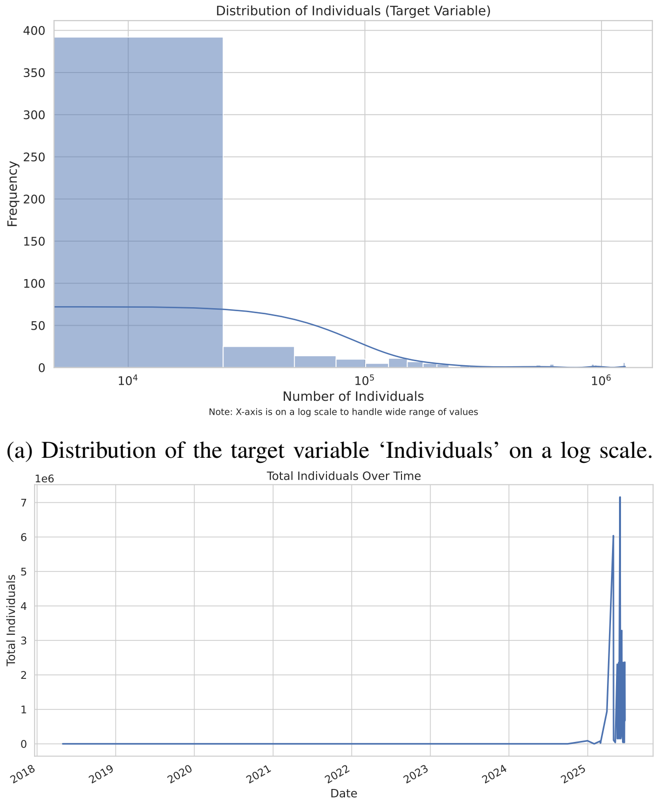
Fig. 2: Distribution of population types in the dataset.

D. Feature Engineering

Three key feature engineering steps were performed to enhance the dataset's predictive power. First, the Date feature was decomposed into separate numerical features to capture potential time-based patterns. For the MLP Regressor, these were Year and Month. For the more granular TabNet model, Year, Month, and DayOfYear were extracted. Second, for the machine learning models, all remaining categorical features, such as 'Country,' 'Population type,' and 'Source', were converted into numerical representations using label encoding. Third, a critical transformation was applied to the highly skewed target variable, Individuals, for both models. A logarithmic function (`np.log1p`) was used to compress the range of the target values. This technique stabilizes the neural network training process and mitigates the influence of extreme outliers, serving as a more robust alternative to manual outlier removal. The models were trained to predict this transformed value, and the final predictions were converted back to their original scale using an exponential function for evaluation.

E. Model Training and Optimization

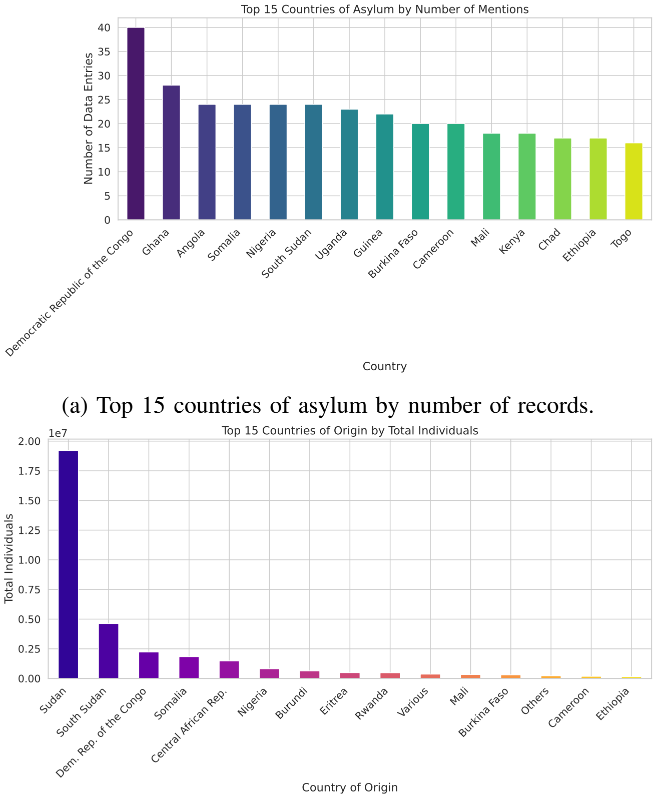
We selected a diverse set of regression models, encompassing both traditional machine learning techniques and advanced deep learning architectures. The traditional models included Decision Tree (DT), k-Nearest Neighbor (kNN), Random Forest (RF), AdaBoost, XGBoost, LightGBM, CatBoost, and



(a) Distribution of the target variable 'Individuals' on a log scale.

(b) Total individuals displaced over time.

Fig. 3: Exploratory Data Analysis of the target variable and its temporal distribution.



(a) Top 15 countries of asylum by number of records.

(b) Top 15 countries of origin by total number of individuals.

Fig. 4: Geographic distribution of displacement events in the dataset.

a Stacking Regressor. Our evaluation also extended to deep learning models, namely a Multi-Layer Perceptron (MLP) Regressor, TabNet [16], FT-Transformer [17], and a pre-trained DistillBERT model [18].

To elicit the best performance from each model, we employed automated hyperparameter tuning. For the traditional machine learning models, we used the 'Randomized-SearchCV' framework. For the deep learning models (MLP Regressor, TabNet, and FT-Transformer), we utilized the Optuna optimization framework, which is specifically designed for efficiently searching large, complex hyperparameter spaces. Both methods aim to find the optimal hyperparameter set θ^* from a defined parameter space Θ by maximizing a cross-validated scoring metric over a fixed number of iterations.

Stacking Ensemble: A two-level Stacking Regressor was implemented to combine the predictive power of multiple strong learners. The architecture, described in Algorithm 1, uses XGBoost, LightGBM, and AdaBoost as level-0 base learners. A Ridge Regressor serves as the level-1 meta-learner, which is trained on the out-of-fold predictions from the base models to produce the final output.

F. Explainable AI (XAI) Framework

The inclusion of an XAI framework is a core part of our methodology, designed to address the "black-box" problem that hinders the adoption of complex ML models in high-stakes applications. By making our model's predictions scrutable, we provide a blueprint for developing more responsible and trustworthy predictive tools. We employed two model-agnostic techniques:

- **LIME (Local Interpretable Model-agnostic Explanations):** LIME explains individual predictions by learning a simpler, interpretable linear model on perturbations of a single data instance [4].
- **SHAP (SHapley Additive exPlanations):** SHAP assigns each feature an importance value for a particular prediction based on principles from cooperative game theory [5]. It explains both local (force plots) and global (summary plots).

This dual approach comprehensively explains the model's micro and macro behavior.

IV. Experiments and Results

A. Setup

The preprocessed dataset was split into training (80%) and testing (20%) sets. All models were trained and tuned on the training set, and their final performance was evaluated on the unseen test set. Performance was measured using standard regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the R-squared (R^2) coefficient.

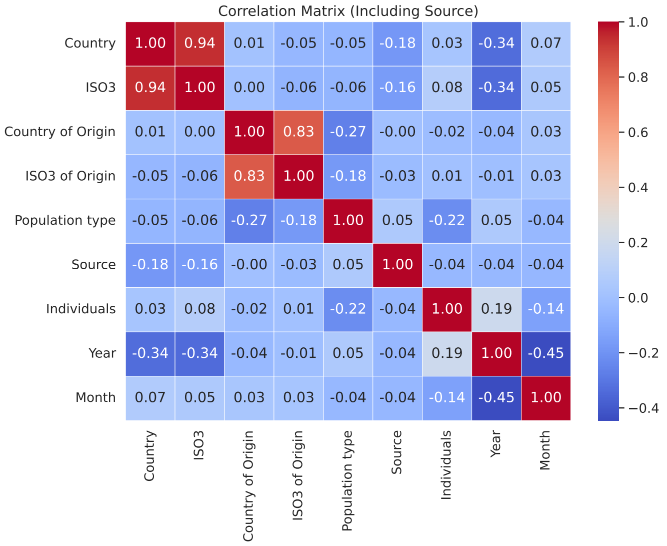


Fig. 5: Correlation matrix of the preprocessed features of the UNHCR dataset.

Algorithm 1 Stacking Regressor Training

```
Let $D_{train} = \{(X_i, y_i)\}_{i=1}^{n}$ be the training set.
Let $M_1, M_2, M_3$ be the base models (XGBoost, LightGBM and AdaBoost), each within a scaling pipeline.
Let $M_{meta}$ be the meta-model (Ridge Regression).
Split $D_{train}$ into $K$ folds for cross-validation (e.g., $K = 5$).
Initialize an empty dataset for meta-features $D_{meta}$.
for $k = 1$ to $K$ do
    Let $D_k$ be the $k$-th fold (hold-out set).
    Let $D_{-k} = D_{train} \setminus D_k$ (training set for base models).
    Train $M_1, M_2, M_3$ on $D_{-k}$.
    Generate predictions $p_{1,k} = M_1(X_k)$, $p_{2,k} = M_2(X_k)$, and $p_{3,k} = M_3(X_k)$ on the hold-out set $D_k$.
    For each instance $X_i \in D_k$, form a new feature vector $[p_{1,i}, p_{2,i}, p_{3,i}]$.
    Append these new vectors to $D_{meta}$ along with their true labels $y_i$.
end for
Train the final meta-model $M_{meta}$ on the complete $D_{meta}$ dataset.
The final Stacking model consists of the base models trained on the full $D_{train}$ and the trained meta-model.
```

The formulas for these metrics are as follows, where y_i is the actual value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of the actual values, and n is the number of samples:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

The experimental setup utilized Python 3.11 with core libraries including Scikit-learn [19] and PyTorch [20].

B. Model Performance

The performance of the applied models with optimized hyperparameters on the unseen test set is presented in Table I. After systematic hyperparameter tuning, the gradient boosting ensemble methods demonstrated markedly superior performance compared to all other models. Our Stacking Regressor model delivered the standout result on the initial test split, achieving an R^2 coefficient of 0.989. XGBoost, LightGBM, CatBoost, AdaBoost, and Random Forest also performed exceptionally well. In contrast, simpler models, including kNN, showed limited predictive power for this dataset, with the MLP Regressor also struggling to capture the complex patterns.

TABLE I: Optimized Model Performance Comparison on Test Set

Model	**MAE**	**MSE**	**RMSE**	**R^2 Coeff.**
DistillBERT	43,011.91	1.69×10^{10}	129,893.23	0.601
TabNet	26,978.16	7.31×10^{9}	85,474.85	0.779
FT-Transformer	22,692.53	5.95×10^{9}	77,138.99	0.820
MLP Regr.	24,052.32	4.81×10^{9}	69,320.75	0.855
kNN	45,843.41	6.85×10^{9}	82,753.79	0.793
DT	9,904.15	1.81×10^{9}	42,570.98	0.945
RF	17,478.27	1.79×10^{9}	42,290.41	0.946
CatBoost	23,178.74	1.35×10^{9}	36,710.71	0.959
AdaBoost	17,254.40	9.07×10^{8}	30,121.37	0.973
LightGBM	24,731.70	1.98×10^{9}	44,445.21	0.940
XGBoost	9,991.01	4.87×10^{8}	22,061.15	0.985
Stacking Ensemble	11,923.29	3.63×10^{8}	19,041.30	**0.989**

C. Cross-Validation for Robustness

To ensure our results were not an artifact of a single “lucky” train-test split comprising the holdout validation approach, we performed 5-fold cross-validation on the top-performing gradient boosting models. The results, summarized in Table II, provide a more robust measure of expected performance. LightGBM emerged as the most consistent and powerful model, with an average R^2 score of 0.88. The low standard deviation of LightGBM’s R^2 score also suggests it may be slightly more stable across different data subsets, compared to models like XGBoost, CatBoost, and AdaBoost. This rigorous validation confirms that our Stacking Regressor was highly overfit.

TABLE II: 5-Fold Cross-Validation Summary for Top Models

Model	Avg. R^2 Score	Avg. RMSE
Stacking Ensemble	0.397 (± 0.207)	15,7920.17 (± 37,102.09)
AdaBoost	0.834 (± 0.175)	71,342.81 (± 50,489.66)
CatBoost	0.823 (± 0.136)	79,979.05 (± 37,042.36)
XGBoost	0.872 (± 0.163)	61,149.50 (± 48,361.59)
LightGBM	**0.879 (± 0.096)**	66,179.91 (± 32,716.86)

TABLE III: Optimized Hyperparameters for the LightGBM Model

Hyperparameter	Value
n_estimators	1083
max_depth	15
num_leaves	88
learning_rate	0.137
subsample	0.732
colsample_bytree	0.937
reg_alpha	83.201
reg_lambda	0.934

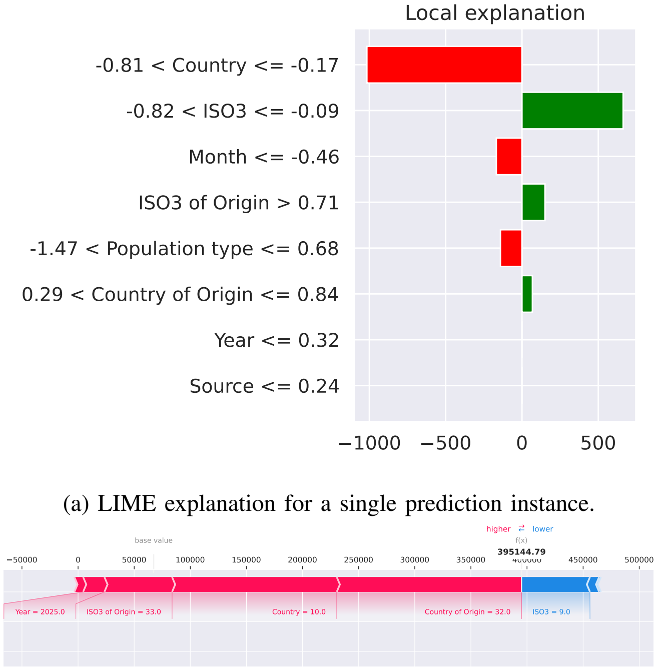
D. Explainable AI Analysis

To interpret the best-performing model (LightGBM, based on the cross-validation technique), we employed both LIME and SHAP. LIME provides local, instance-level explanations. For one sample prediction, LIME identified that the country of origin and asylum were the primary drivers, depicted in Fig. 6a. SHAP provides both local and global explanations. The SHAP force plot in Fig. 6b offers a more detailed local view, showing how each feature value pushes the prediction higher or lower than the baseline.

Globally, the SHAP summary plots in Fig. 7 provide a comprehensive view of feature importance. The bar plot (Fig. 7a) shows the mean absolute SHAP value for each feature, clearly indicating that 'Country of Origin', 'ISO3' (country of asylum), and 'Country' are the top three most influential predictors, followed by 'Year'. The dot plot (Fig. 7b) offers deeper insight, revealing not only the magnitude of a feature's impact but also its direction. For instance, it shows that high values for 'Year' (more recent events) tend to have a positive impact on the predicted number of individuals. Together, these plots confirm that the geographic dyad and the time of the event are the most critical predictors in the model.

E. Comparison with Prior Work

To contextualize our findings, Table IV benchmarks our results against prior works in migration forecasting. While direct comparisons are challenging due to different datasets and specific objectives, this table highlights that our approach achieves state-of-the-art performance within this research domain. The high R^2 value demonstrates the effectiveness of our optimized ensemble and XAI framework on a recent, relevant dataset. The diverse range of datasets and models used in previous research, from satellite imagery with Mask R-CNN to ensemble methods with Google Trends, underscores the varied approaches in refugee and migration forecasting.



(a) LIME explanation for a single prediction instance.

(b) SHAP force plot for the same instance, showing feature impacts.

Fig. 6: Local, instance-level explanations from LIME and SHAP.

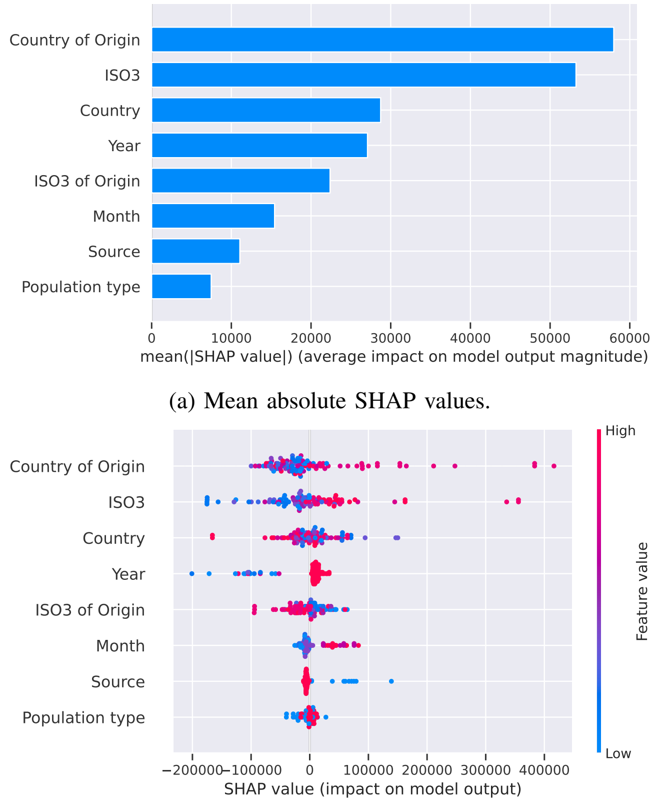
F. Discussion

The strong cross-validated performance of the tuned LightGBM model affirms its suitability for complex socio-political prediction tasks. The model's success lies in capturing the intricate, non-linear interactions between geographical and temporal factors that drive migration. The application of our XAI framework moves this research beyond a simple performance comparison. We validate the model's logic using LIME and SHAP to provide the granular insights necessary for practical application and stakeholder trust. This phenomenon opens avenues for causal inference and allows researchers to form hypotheses about the key drivers of specific migration events.

This model would serve as a decision-support tool rather than an automated system for a real-world deployment. Real-world validation would involve forecasting future flows and comparing predictions against incoming UNHCR data, allowing continuous model evaluation. A key consideration would be model drift; the system would require periodic retraining on new data to adapt to evolving geopolitical landscapes. Integrating these predictions and their corresponding SHAP explanations into a dashboard would empower humanitarian planners to see a forecast and understand the factors driving it, leading to more informed and defensible resource allocation decisions.

TABLE IV: Comparison of the Proposed Refugee Prediction System with Prior Related Works

Reference	Dataset	Sample Count	Best Model	Result
Boss et al. [1]	EU Asylum, Google Trends	N/A	Ensemble (RF, XGB)	Outperformed RW benchmark
De los Santos et al. [2]	UNHCR, GDELT	N/A	Grad. Boost	20% lower relative RMSE
Carammia et al. [6]	EASO, GDELT	N/A	Dynamic Elastic Net	7% avg. relative error
Quinn et al. [21]	Satellite Imagery	87k structs.	Mask R-CNN	0.78 Mean Avg. Precision
Bosco et al. [22]	Frontex, FAO	169 (monthly)	Ensemble (ANN, RF, XGB)	R^2 = 0.90 (validation)
Pham & Luengo-Oroz [13]	UNHCR PRMN (Somalia)	N/A	MLP	RMSE = 6,288
Chen & Eagel [23]	US Asylum Hearings	492k+	Random Forest	82% Accuracy
This Study	**UNHCR Situations**	**506 (updates monthly)**	**Stacking Ensemble**	**0.989 $\boldsymbol{R^2}$**



(a) Mean absolute SHAP values.

(b) SHAP value distribution.

Fig. 7: SHAP global summary plots illustrating feature importance.

Limitations: This study has several limitations. First, our model relies solely on historical UNHCR data and does not incorporate external, real-time data sources, including conflict event data or economic indicators [1], [2]. Second, while necessary for model stability, the outlier removal process means our model may be less accurate at predicting rare, large-scale crises.

G. Ethical Considerations

The deployment of predictive models in humanitarian contexts raises critical ethical questions. While our approach aims to improve resource allocation, care must be taken to ensure predictions do not reinforce existing biases or influence restrictive immigration policies. We recommend a human-in-the-loop approach where domain experts validate model outputs before policy decisions [24].

V. Conclusion

This research successfully developed and evaluated a machine learning framework for predicting the scale of global refugee and asylum seeker flows. By systematically comparing various models and employing rigorous hyperparameter tuning and cross-validation, we demonstrated that optimized gradient boosting models, particularly LightGBM and XGBoost, provide state-of-the-art predictive accuracy. Crucially, our methodological emphasis on integrating Explainable AI through LIME and SHAP allowed us to move beyond black-box predictions to interpret the key factors influencing displacement forecasts. This capability is vital for building trust and providing actionable intelligence to humanitarian organizations.

Future Work: Future research should focus on three key areas. First, enriching the feature space by integrating dynamic, external data sources will be critical for moving from historical analysis to accurate forecasting. Second, advanced techniques for handling skewed data and modeling extreme events, such as quantile regression, should be explored. Finally, a deeper analysis of the temporal dynamics using time-series-specific models could further improve predictive power.

References

[1] K. Boss, A. Groeger, T. Heidland, F. Krueger, and C. Zheng, "Forecasting bilateral asylum seeker flows with high-dimensional data and machine learning techniques," *Journal of Economic Geography*, vol. 25, pp. 3–19, 2025.

[2] D. De los Santos, E. Frey, and R. Vassallo, "Forecasting global refugee flows: a machine learning approach using non-conventional data," Master's thesis, Barcelona School of Economics, 2023.

[3] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, 2021.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" explaining the predictions of any classifier," in *International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[6] M. Carammia, S. M. Iacus, and T. Wilkin, "Forecasting asylum-related migration flows with machine learning and data at scale," *Scientific Reports*, vol. 12, 2022.

[7] C. Havas, L. Wendlinger, J. Stier, S. Julka, V. Krieger, C. Ferner, A. Petutschnig, M. Granitzer, S. Wegenkittl, and B. Resch, "Spatio-temporal machine learning analysis of social media data and refugee movement statistics," *International Journal of Geo-Information*, vol. 10, 2021.

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[9] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–259, 1992.

[10] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[11] A. Beduschi, "International migration management in the age of artificial intelligence," *Migration Studies*, vol. 9, pp. 576–596, 2021.

[12] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information fusion*, vol. 58, pp. 82–115, 2020.

[13] K. Hoffmann Pham and M. Luengo-Oroz, "Predictive modelling of movements of refugees and internally displaced people: towards a computational framework," *Journal of Ethnic and Migration Studies*, vol. 49, pp. 408–444, 2023.

[14] N. Kinchin, "The human in the feedback loop: Predictive analytics in refugee status determination," *Law, Technology and Humans*, vol. 6, pp. 23–45, 2024.

[15] UNHCR, "UNHCR situations - refugee data." https://data.humdata.org/dataset/unhcr-situations, 2024.

[16] S. Ö. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 6679–6687, 2021.

[17] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," *Advances in neural information processing systems*, vol. 34, pp. 18932–18943, 2021.

[18] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[20] A. Paszke, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[21] J. A. Quinn, M. M. Nyhan, C. Navarro, D. Coluccia, L. Bromley, and M. Luengo-Oroz, "Humanitarian applications of machine learning with remote-sensing data: review and case study in refugee settlement mapping," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, 2018.

[22] C. Bosco, U. Minora, A. Rosińska, M. Teobaldelli, and M. Belmonte, "A machine learning architecture to forecast irregular border crossings and asylum requests for policy support in Europe: a case study," *Data & Policy*, vol. 6, 2024.

[23] D. L. Chen and J. Eagel, "Can machine learning help predict the outcome of asylum adjudications?," in *International Conference on Artificial Intelligence and Law*, pp. 237–240, 2017.

[24] M. DeDonato, V. Dimitrov, R. Du, R. Giovacchini, K. Knoedler, X. Long, F. Polido, M. A. Gennert, T. Padır, S. Feng, *et al.*, "Human-in-the-loop control of a humanoid robot for disaster response: a report from the DARPA robotics challenge trials," *Journal of Field Robotics*, vol. 32, pp. 275–292, 2015.