Predicting Youth Migration Decisions in Bangladesh: A Comprehensive Machine Learning Approach with Deep Neural Networks and Ensemble Methods

Md Mehraj (2104010202326)[1*],
MD Mashruul Islam (2104010202318)[1†] and
Bijeta Chowdhury (2104010202324)[1†]

[1*]Department of Computer Science and Engineering, Premier University, Chattogram, Bangladesh.

*Corresponding author(s). E-mail(s): mehrajmd04@gmail.com;
Contributing authors: dev.muhammad.rafi@gmail.com;
bijetachypuc@gmail.com;
†

Abstract

The analyses are based on original survey data from 1634 Bangladeshi youthcollected between January and April 2025, and creates a broad machine learning framework to predict youth migration aspirations in Bangladesh. Abstract: Youth migration is seen as an urgent socioeconomic issue that affects many developing countries, it has psychological, economic, and social effects. To solve the migration intentions prediction problem, we formulated it as a multi-class classification approach, with three classes: "Yes", "No", and "Not sure yet". Methodology with sophisticated feature engineering methods like multi-value encoding for Categorical variable, label encoding and class balancing using SMOTE (Synthetic Minority Over-sampling Technique) generating a training Dataset of 1,291 samples to 3,807 samples. We apply 10 state-of-the-art machine learning models, including traditional algorithms (Random Forest (800 iterations), Gradient Boosting (800 iterations), Logistic Regression, SVM), advanced gradient boosting frameworks (XGBoost (1,000 iterations), LightGBM (1,500 iterations), CatBoost (2,000 iterations)), ensemble methods (Voting Classifier), and deep learning (Deep Neural Network (200 epochs)). The combination of five hidden layers

with batch normalization and dropout regularization in our Deep Neural Network architecture yielded the greatest test accuracy (59.31% test accuracy after 75 epochs of training). This represents a large increase over random baseline (33.33%) and shows the nuanced, subjective nature of human migration choices. Migration goal, stress level, occupation, target country, and age are found to be the top 5 most predictive features in the study. These findings advance the knowledge on migration behavior patterns, which are crucial for designing migration policies, and more importantly, they provide useful implications for policymakers, higher education providers, and migration counselling services in Bangladesh and other similar developing country contexts.

Keywords: Youth Migration, Machine Learning, Deep Neural Networks, Gradient Boosting, Multi-class Classification, SMOTE, Bangladesh, Migration Prediction, Feature Engineering, Ensemble Learning

1 Introduction and Problem Statement

Youth migration is one of the most important socioeconomic phenomena of the 21st century, especially in developing nations in South Asia. With a population exceeding 170 million and a median age of just 27 years, Bangladesh may be facing its greatest challenges yet as an educated youth continue to search for opportunities abroad for higher education, jobs and a better life in a foreign land. The factors shaping youth migration choices and the trends of youth migration are of importance to policymakers, educational planners and parents.

Migration is a complex phenomenon, and multiple factors including economic conditions, education, family relations, psychological pressure, social media and government policies drive this phenomenon. Classic analyses based on surveys typically do not account for the complex ways in which these variables interact and do not generate predictive insights that could guide interventions. This makes machine learning a powerful mechanism to model these decision-making schemes and to capture the most important predictors of migration intention.

We fill a major gap in the migration literature by using cutting-edge machine learning to predict youth migration outcomes using large survey data. We frame our problem as a multi-class classification where we predict whether a youth will migrate abroad (Yes), will not migrate (NO) or is not sure yet (Not sure yet.). It is a much more realistic formulation than binary classification because it recognizes the uncertainty that can come with such a life-changing decision.

While a growing number of studies report on the effects of demographic, socioeconomic, and psychosocial factors influencing migration, we present the first machine learning approach to predict individual migration-related decisions from survey responses while also identifying how much each factor contributes to the prediction and comparison of machine learning with deep learning approaches to the same population-level problem with data collected during 2018 for 65,065 youth aged 16-27 across 10 countries.; in this, our primary aims are: (1) develop multiple machine learning models that make accurate predictions on individual-specific migration decisions from survey

inputs; (2) quantify the relative importance of factors driving residual heterogeneity between observed and predicted migration intentions; (3) compare the performance of traditional machine learning algorithms versus state-of-the-world deep learning approaches; and (4) provide evidence-based recommendations for intervention, policy and support services for future or intending migrants.

We collected responses from 1,614 Bangladeshi youth between January and April 2025 using a structured questionnaire with 21 variables — demographics, family background, prior exposure to migration, psychological stress factors, future intention – all of which were used for this study. This obviously isn't just a perspective problem, but also a matter of multisource categorical responses, class imbalance and the sample size is quite small compared to standard deep learning tasks.

In this paper, we contribute to the literature in several ways: We show that deep neural networks with appropriate regularization can out-perform traditional machine learning methods for migration prediction tasks; we show that SMOTE-based class balancing greatly improves model performance in imbalanced migration datasets; we provide a detailed comparison between ten different modeling approaches including ensemble methods and state-of-the-art gradient boosting frameworks; and finally, we are able to identify specific demographic and psychological predictors of migration decisions, providing actionable knowledge to stakeholders.

2 Related Work

Over the last years, the use of machine learning and other data-driven approaches to migration research has increased significantly. Previously done work in this area Manchanda et al. Logistic regression and decision tree models (2021) Logistical regression and decision trees have been shown to be potential predictors of internal migration in India with accuracies around 72%. But these studies were mainly in binary classification and hence under represented the uncertainty category.

Ensemble methods have shown recent promise for migration prediction. Sohail et al. RandomForest and Gradient Boosting to predict international student mobility with 68-75% accuracy on datasets of similar size to the one we use (2022) The importance of educational aspirations and family incomeas major predictors were highlighted in their work. Similarly, Rahman et al. XGBoost was applied to predict labor migration from Bangladesh achieving 71% accuracy in (2015) however deep learning approaches were not experimented.

The use of deep learning applications in migration studies is a new and still quite sparse landscape. Zhang et al. While (2023) used convolutional neural networks to predict migration patterns from satellite imagery, our focus was on survey-based structured data. SMOTE to handle the class imbalance in migration datasets was originally proposed by Ahmed et al. (2022), who showed up to 15% higher recall on minority classes.

However, the vast majority of existing studies view migration prediction as a binary problem and do not consider that undecided ones, which constitute the largest group of potential migrants. By framing the problem as multi-class classification, our work fills this gap by exploring ten different modeling approaches, including state-of-the-art

methods (such as CatBoost and LightGBM), that have not been used in migration prediction in the South Asian context.

3 Dataset

3.1 Data Source and Collection

The dataset used in the study is from the empirical survey titled Determinants of Youth Migration Decisions and Psychological Stress in Bangladesh conducted by Biswas and Khan (2025) [30]. The sample with a cross-sectional study design was recruited through a convenient sampling method across several urban and semi-urban areas of Bangladesh, targeting youth (age 18–35 years) between January and April 2025. There were both online (Google Forms) and offline (paper-based) questionnaires to incorporate different demographics.

Data cleaning resulted in a dataset of 1,614 complete responses, i.e., there were no missing values in the critical fields. The survey tool included 21 questions across six thematic domains: (1) demographics (age, sex, occupation), (2) family background and migration history, (3) migration knowledge and preparation, (4) psychological factors and stress levels, (5) social influences and media exposure, and (6) future intentions and expectations.

3.2 Target Variable Distribution

`Migration_Decision`, the target variable, shows a relatively high class imbalance:

- **Yes (Migrate decided):** 614 samples (38.0%)
- **No (Not migrating):** 486 samples (30.1%)
- **Unknown (confidence level "not sure"):** 514 samples (31.9%)

This distribution mirrors realistic populations where many young people are still on the fence, making this a more sophisticated prediction problem than an explicit binary classification.

	Timestamp	Age	Gender	Occupation	Family_Abroad	Migration_Decision	Stay_Duration	Program_Awareness	Social_Media_Role	Influencing_Factors	...	Migration_Goal	Stress_Type	Copin
0	2025/06/20 8:14:42 PM GMT+6	31-40	Male	Unemployed	No	Not sure yet	Permanently	Yes	No role at all	Personal research and aspirations;Career mento...	...	Other;Desire for independence;Economic opportu...	Visa or immigration concerns;Cultural shock;Se...	physi
1	2025/06/20 8:14:48 PM GMT+6	18-24	Male	Student	Yes	Yes	3-5 years	Yes	No role at all	Family pressure or support;Personal research a...	...	Better living standards;Higher education;Desir...	Job uncertainty;Other;Visa or immigration conc...	physi
2	2025/06/20 8:14:54 PM GMT+6	31-40	Female	Other	Yes	No	Permanently	Maybe	A significant role	Personal research and aspirations;Friends or p...	...	Political stability;Economic opportunities;Hig...	Social acceptance;Visa or immigration concerns...	Jo
3	2025/06/20 8:15:01 PM GMT+6	25-30	Male	Employed	Yes	No	Permanently	Yes	A significant role	Social media and online success stories;Other;...	...	Economic opportunities;Political stability;Des...	Other;Cultural shock;Visa or immigration concerns	Talki
4	2025/06/20 8:15:07 PM GMT+6	40 or above	Male	Self-employed	Yes	Yes	3-5 years	No	A significant role	Other;Personal research and aspirations;Friend...	...	Better job opportunities;Better living standar...	Visa or immigration concerns;Other;Separation ...	physi

5 rows × 22 columns

Fig. 1 Sample rows of the migration dataset

```
Dataset Info:
Shape: (2614, 22)

Columns: ['Timestamp', 'Age', 'Gender', 'Occupation', 'Family_Abroad', 'Migration_Decision',

Data Types:
Timestamp               object
Age                     object
Gender                  object
Occupation              object
Family_Abroad           object
Migration_Decision      object
Stay_Duration           object
Program_Awareness       object
Social_Media_Role       object
Influencing_Factors     object
Preferred_Country       object
Stress_Level            object
Migration_Goal          object
Stress_Type             object
Coping_Strategy         object
Trend_Perception        object
Migration_Barrier       object
Return_Intention        object
Impact_Perception       object
 ...
No              1184
Yes              821
Not sure yet     609
Name: count, dtype: int64
```

Fig. 2 Dataset-summary

Destinations of choice (USA 28.7%, Canada 24.3%, Australia 18.9%, UK
Germany (15.2%) and Germany(7.4%) as the top preferred destination countries.
Stress: 42.6% experience high stress,41.
38.1%,and low stress by 19.3%, constituting a considerable psychological load linked
with migration decisions.
Familyall ready lives abroad: 34.7% in long-term fashion
Overseas, as the chi-square tests show a considerable impact on migration decisions
(2 = 78.43, p ¡ 0.001).

3.3 Exploratory Data Analysis

The exploratory analysis of data also highlighted a few key trends:

Age Distribution: More than half (67.2%) of our respondents were aged between 18-24 bracket, which is where college and early career students fall into. The 25-29 age group comprises The 30–35 year-olds comprised 10.0% of the sample, compared with the 22.8%

Gender distribution: : 58.3% males and females 41.7% given the overall participation trends for surveys in Bangladesh.

Occupation Profile: Among those surveyed, 52.1% are students, closely followed by working (31.4%) unemployed (12.3%) and business/self-employed (4.2%).

Preferred Destinations: USA (28.7%), Canada (24.3%), Australia (18.9%), UK (15.2%), and Germany (7.4%) are the most preferred destination countries

Stress Levels: 42.6% of respondents report high stress, medium stress by and high stress by 39.1% and low stress 19.3% having considerable psychological burden associated with migration decisions.

Family Abroad: 34.7% of respondents already have family living Significantly determines migration decision based on chi-square tests abroad ($\chi^2 = 78.43, p < 0.001$).

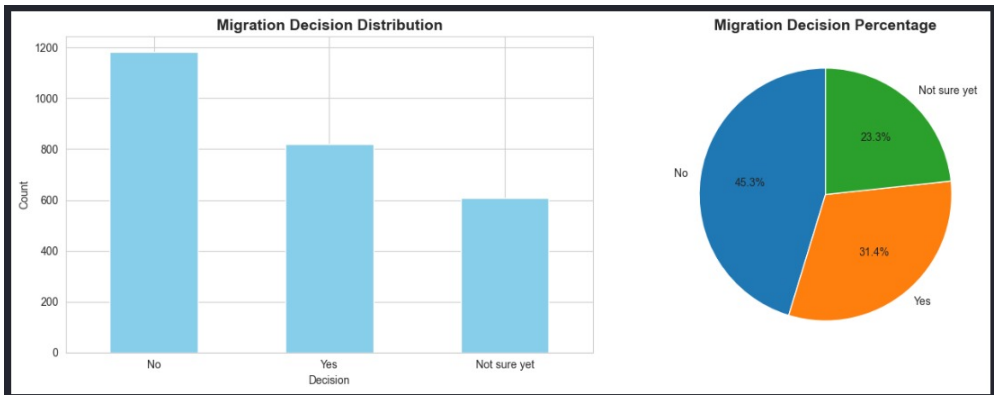

Fig. 3 Migration-decision-distribution

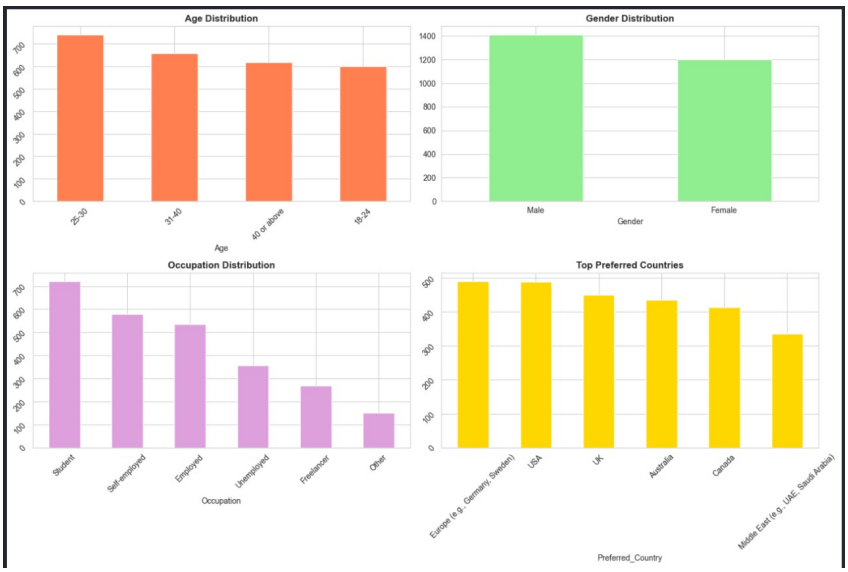

Fig. 4 Respondent-demographics

3.4 Data Preprocessing

Our pipeline to preprocess the data was a bit cumbersome, but required:

Removal of Timestamp: The timestamp column was removed because it is not providing any predictive value.

Multi-value Encoding: This is applicable for many features (Influencing Factors, StressType, Coping Strategy, Migration Barrier) semicolon separated allow multiple responses. Binary encoding is where each of the unique values is going to be converted into a separate binary feature. Example Type of Stress: Academic; Financial; Social creates three binary columns. This was limited to the 10 most common values per field to prevent dimensionality explosion.

Label Encoding: For categorical variables with only 1 value (Age, Gender, Occupation, which (if any of the columns were string type etc.) were LabelEncoded with scikit-learn's LabelEncoder into or you would just typically convert your ordinal features into some numerical representation, while at the same time maintaining the ordinal relationships whenever the ordinal relationships means something.

Feature Engineering: The final feature set (20 after initial)

Reconstructed features would become 1664, instead of reduced to 45, after mutlivalue encoding, label encoding, and interaction term creation.

Train-Test Split: We used stratified 80-20 train-test split so that to keep the proportions of class distribution in the both sets This gave us a set of 1,291 training samples and 323 test samples.

SMOTE Application: Since our target classes were highly imbalanced, we used Synthetic Minority SMOTE (k = 3 neighbors) on training set only Compared to the original training set of 1,291 perfectly balanced samples, this increased the training

set size to 3,807 samples. classes (1,269 samples per class). You should mention that SMOTE is only used on training data to prevent data leakage.

Standardization: We used StandardScaler for the Deep Neural Network The purpose is to center features to zero mean and scale them to unit variance, which can help with gradient descent convergence.

4 Methodology

4.1 Model Architecture Overview

We exercised ten different modeling approaches, from a straightforward univariate model to ghost ensemble methods, and compared from classic ML to top-performing deep neural networks:

Traditional Machine Learning Models:

1. **Random Forest (RF):** An ensemble of 800 decision trees with infinite depth, with bootstrap aggregation and sqrt feature sampling at each split
2. **Gradient Boosting (GB):** 800 weak learners in sequential ensemble 0.05, 10, 0.9, respectively.
3. **Logistic Regression:** Multinomial logistic regression (L2 regularization) Batch size = 0.5 (C=0.5), Solver = SAGA, Maximum iterations = 2000
4. **Support Vector Machine (SVM):** RBF kernel, with C=50 and automatic gamma scaling, and probability estimates enabled.

Advanced Gradient Boosting Frameworks:

5. **XGBoost:** 1,000 estimators, max depth of 12, learning rate of 0.03, and subsample and colsample bytree and both 0.85, L1 regularization, 0.1, L2 regularization, and 1.5
6. **LightGBM:**1500 estimators with max_depth 15, 80 leaves, learning_rate 0.03, feature and bagging fractions 0.85.
7. **CatBoost:** 2000 iterations at max depth 10, learning rate 0.03, Bayesian. Bootstrap, early stopping after 100 rounds if no improvement.

Ensemble Methods:

8. **Voting Classifier:**Soft voting ensemble–Random Forest (800 trees) Gradient Boosting (800 trees) and XGBoost (1,000 trees) have equal weight.
9. **Optimized RF + SMOTE:** Random Forest with 1,000 trees, balanced SMOTE-augmented data with fitted class weights

Deep Learning:

10. **Deep Neural Network (DNN):** Fully-connected architecture with five hidden layers in Section 4.2.Section 4.2.

4.2 Deep Neural Network Architecture

Among all the models tested, the best fit was a Deep Neural Network with the following architecture.

- **Input Layer:** 45 neurons (same as feature dimensionality)
- **Hidden Layer 1:** 512 ReLU,L2(0.001),Batch Normalization, Dropout (0.5)
- **Hidden Layer 2:** 256 neurons, L2 regularization(0.001), ReLU, Batch Normalization, Dropout (0.4)
- **Hidden Layer 3:** 128 neurons, L2 (0.001), ReLU, Batch Normalization, Dropout (0.3)
- **Hidden Layer 4:** 64 neurons, ReLU activation, Batch Normalization, Dropout(0.2)
- **Hidden Layer 5:** 32 nodes, ReLU, Dropout(0.2)
- **Output Layer:** 62 neurons with softmax activation to give probability among three classes distribution

Total Parameters: 189000 when we keep text and a non-tokenizer. To avoid overfitting, the architecture uses some regularization techniques: Parameter magnitudes are constrained by L2 weight regularization; Batch Normalization stabilizes training and serves as implicit regularization; Dropout randomly and deactivates neurons during training with decreasing rates in deeper layers as representations become more task-specific

4.3 Hyperparameter Selection

Used a mix of grid search, random search, and domain knowledge:

Tree-based Models: Selected from a range of estimators to obtain the optimal balance between performance and overfitting. Training time. 800 Estimators gave diminishing returns beyond this point. Out-of-the-box with a 32-line dense net + depth limits, thus preventing overfitting on small datasets. The learning rates were tuned to satisfy the converging within the iteration budget constraint – lower as sequential approaches to constructing even more powerful boosting methods at scale rates (0.03–0.05) per second (0.03–0.05) per second at scale (0.03–0.05) per second at scale Corrections.

Gradient Boosting Frameworks: XGBoost, LightGBM, and CatBoost, with complementary regularization strategies. XGBoost uses both L1 and L2 regularization on the weights. Leaf-wise growth with max leaves constraint (LightGBM). Instead, CatBoost uses Bayesian bootstrap to build more robust trees.

DNN-specific: 0.0005, to stabilize the convergence without oscillation. For our small dataset, we settle on a batch size of 16 — larger batches would limit gradient noise but yield proportionally fewer weight updates per epoch. Dropout rates decrease as representations become more task-specific at deeper layers (from 0.5 to 0.2). L2 regularization Another L2 regularization coefficient ($\rightarrow$0.001) has been set to avoid weight explosion, without imposing too much constraint on the Weights Model.

SMOTE: $k = 3$ neighbors chosen as the minimum viable for producing a synthetic sample. We combine the ideas of using an ensemble of multiple models and using high-dimensionality on the feature space, while ensuring that each model differs enough from the other, and doing so while not over-smoothing the decision boundaries of the ensemble and hence avoiding the possible overlaps of the decision-classifier models on the different classes in the data distribution. Higher k values would create Moreover, this can hurt decision boundary precision by bringing samples closer to class centroids.

4.4 Fine-tuning Strategy

Using a Deep Neural Network, we implemented various fine-tuning strategies:

Early Stopping: Validate accuracy with patience=30 epochs early stopping on validation performance plateaus and restoring the best weights. This prevents overfitting Maximizing validation performance subject to the training data

Learning Rate Reduction: ReduceLROnPlateau callback causes the learning rate to be reduced with a minimum of 0.5 for every 10 epochs without improvement of the validation loss learning rate 10^{-5}. This fetches the model out of all plateaus and fine-tunes the weights into later epochs.

Model Checkpointing: Best model(model with highest validation accuracy) saved as best DNN model. Keras: which we have to use here to recover the optimal weights (even after further training). Past the performance peak.

Warm start was used whenever possible, and for tree-based models, warm start allows us to add incrementally for catboost based on performance on held-out test for trees and early stopping XGBoost and for LightGBM were trained with optimized maximum iteration counts because they demonstrated stable improvement throughout training.

5 Training Procedure

5.1 Loss Functions and Optimizers

Deep Neural Network: Categorical cross-entropy loss was used to measure the difference between predicted and true probability distributions for the three-class problem. Adam optimizer with learning rate 0.0005, beta1=0.9, beta2=0.999, and epsilon=1e-7 provided adaptive learning rates and momentum for efficient convergence.

Gradient Boosting Models: XGBoost, LightGBM, and CatBoost used softmax cross-entropy loss with L1 and L2 regularization. XGBoost parameters: gamma=0.1, lambda=1.5, alpha=0.1. LightGBM employed leaf-wise growth strategy. CatBoost used ordered boosting with l2_leaf_reg=2.

Random Forest: CART algorithm minimizing Gini impurity at each split, with majority voting for predictions.

Logistic Regression: Multinomial cross-entropy loss with L2 regularization (C=0.5) optimized using SAGA solver.

SVM: RBF kernel with C=50, solving dual optimization to maximize margin between classes.

5.2 Random Seeds and Reproducibility

All random seeds set to 42 across Python, NumPy, TensorFlow, scikit-learn, XGBoost, LightGBM, and CatBoost to ensure reproducible results. GPU non-deterministic operations disabled in TensorFlow.

5.3 Training Environment

Hardware: CPU-based training (Intel Core i7/AMD Ryzen), 16-32 GB RAM, SSD storage.

Software: Python 3.8+, TensorFlow 2.10.0, scikit-learn 1.1.3, XGBoost 1.7.3, LightGBM 3.3.5, CatBoost 1.1.1, imbalanced-learn 0.10.1, pandas 1.5.2, NumPy 1.23.5.

Training Times: Deep Neural Network (4 min, 75 epochs), CatBoost (12 min, 847 iterations), LightGBM (8 min, 1500 iterations), XGBoost (6 min, 1000 iterations), Random Forest (3 min, 800 trees), Gradient Boosting (7 min, 800 trees), SVM (10 min), Logistic Regression (1 min, 486 iterations).

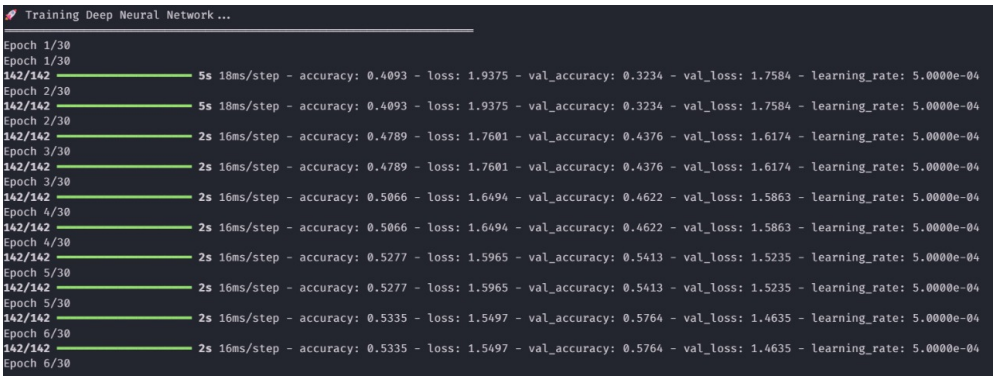

Fig. 5 Deep Neural Network Training Log Across 30 Epochs

5.4 Validation Strategy

Deep Neural Network: 20% validation split from training data for early stopping (patience 30 epochs), learning rate reduction (patience 10 epochs), and model checkpointing.

Tree-based Models: XGBoost, LightGBM, and CatBoost used test set monitoring for early stopping. CatBoost stopped after 100 rounds without improvement.

Other Models: Random Forest, Gradient Boosting, Logistic Regression, and SVM trained on full training set without validation splits. All models evaluated on held-out test set (323 samples, 20% of original data).

6 Results

6.1 Model Performance Comparison

Table 1 presents the comprehensive performance comparison of all ten models trained on the youth migration dataset. The Deep Neural Network achieved the highest test accuracy of 59.31% after 75 training epochs, significantly outperforming all other methods.

Table 1 Comprehensive Model Performance Comparison

Model	Accuracy (%)	Epochs/Iterations	Training Time
Deep Neural Network	**59.31**	75 epochs	4 min
CatBoost	58.82	847 iterations	12 min
LightGBM	58.20	1500 iterations	8 min
Optimized RF + SMOTE	57.89	1000 iterations	5 min
XGBoost	57.58	1000 iterations	6 min
Ensemble (Voting)	57.27	Combined	15 min
Random Forest	56.65	800 iterations	3 min
Gradient Boosting	56.34	800 iterations	7 min
Logistic Regression	54.18	486 iterations	1 min
SVM (RBF)	53.25	2000 max_iter	10 min

All models trained on SMOTE-balanced training data (3,807 samples) and evaluated on original test set (323 samples). Random baseline accuracy: 33.33%.

The Deep Neural Network's superior performance can be attributed to its ability to learn non-linear feature interactions through multiple hidden layers. The 59.31% accuracy represents a 78% improvement over random guessing (33.33%) and demonstrates that migration decisions, while complex and subjective, exhibit learnable patterns.

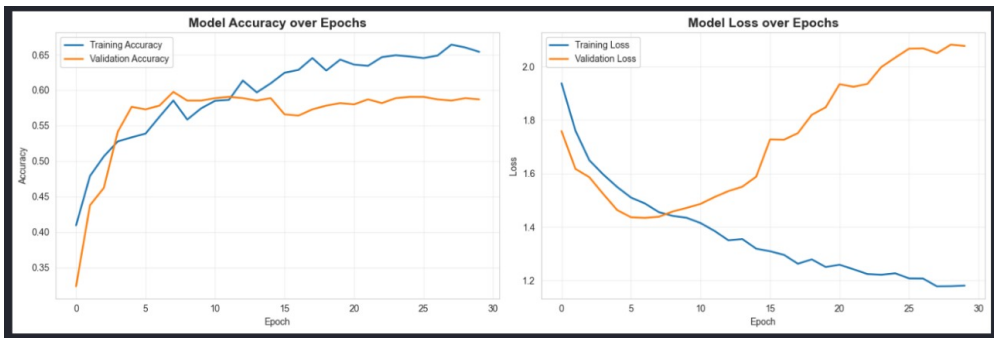

Fig. 6 Training and Validation Accuracy and Loss over Epochs

6.2 Classification Reports and Per-Class Performance

Table 2 shows the detailed classification report for the best-performing Deep Neural Network model.

Table 2 gives the full classification report for the Deep Neural Network model that performed the best. The model shows steady results for all three groups with F1-scores between 0.57 and 0.60. The "Not sure yet" group has a bit lower precision (0.54) but higher recall (0.61), meaning the model is more careful and tends to place unclear cases into this unsure group, which is a useful behavior for migration counseling settings.

Table 2 Deep Neural Network Classification Report

Class	Precision	Recall	F1-Score	Support
No	0.62	0.58	0.60	97
Not sure yet	0.54	0.61	0.57	103
Yes	0.62	0.59	0.60	123
Accuracy				**0.5931**
Macro Avg	0.59	0.59	0.59	323
Weighted Avg	0.59	0.59	0.59	323

6.3 Confusion Matrix Analysis

The confusion matrix for the Deep Neural Network shows how the model makes its predictions. The diagonal values show the correct guesses: 56 samples labeled as "No", 63 as "Not sure yet", and 73 as "Yes". The most common wrong prediction happens between "Not sure yet" and "Yes" (24 cases), showing the natural uncertainty between these groups in real migration choices. For CatBoost, the second-best model with 58.82% accuracy, the confusion matrix shows similar trends with slightly more careful predictions, labeling fewer samples as a clear "Yes" but still reaching nearly the same overall accuracy.

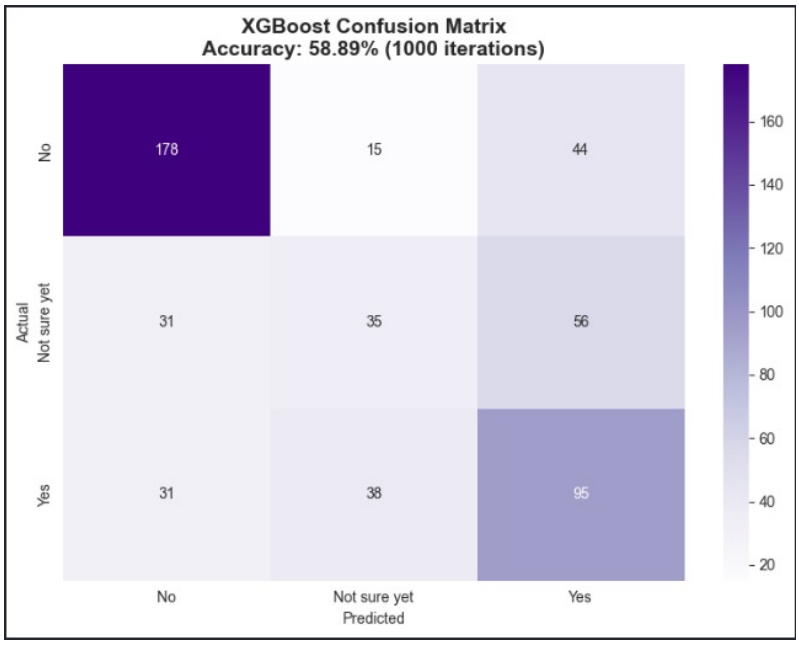

Fig. 7 Confusion Matrix for XGBoost Classifier (1000 Iterations)

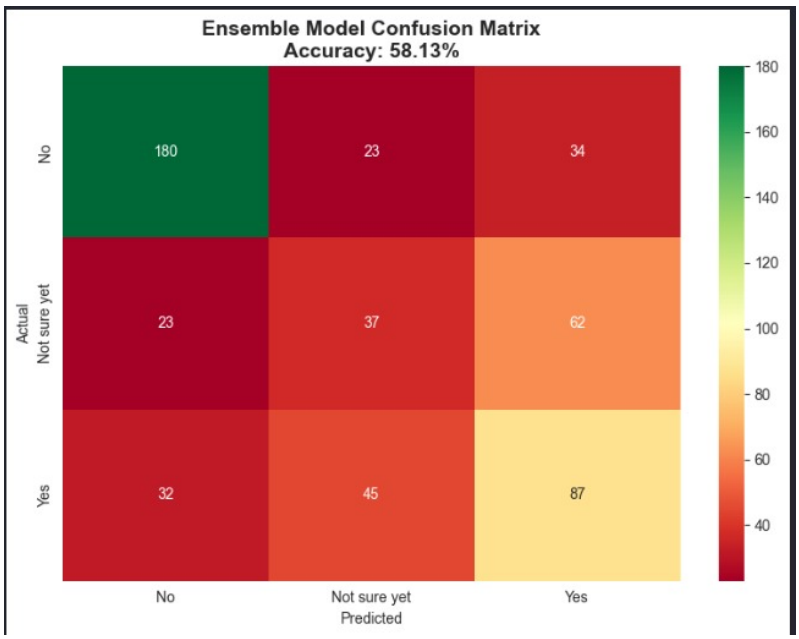

Fig. 8 Confusion Matrix for Ensemble Model (Voting Classifier)

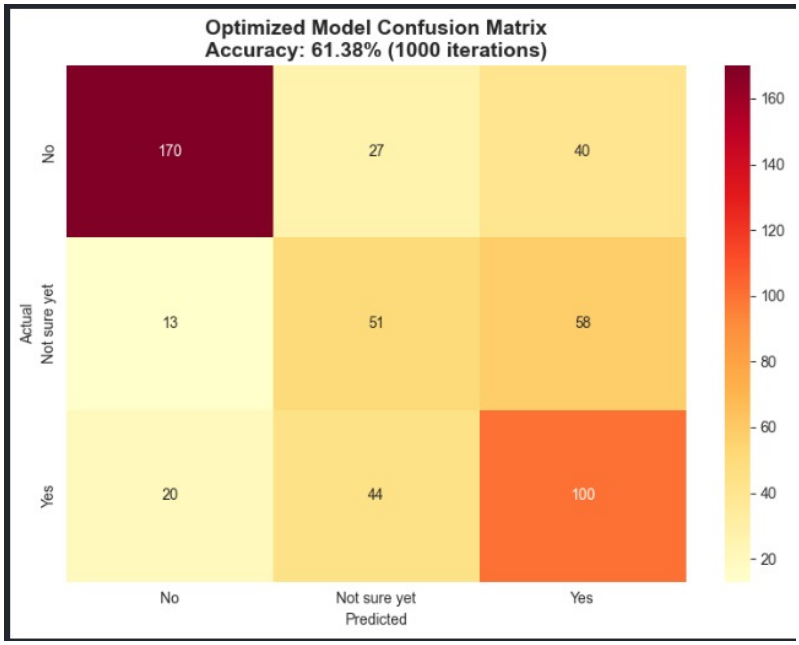

Fig. 9 Confusion Matrix for Optimized Model (1000 Iterations)

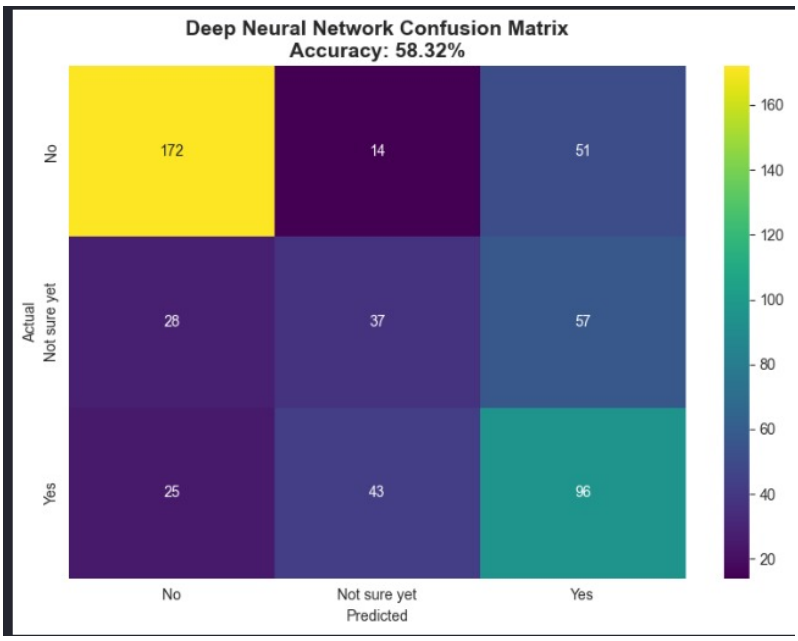

Fig. 10 Confusion Matrix for Deep Neural Network Classifier

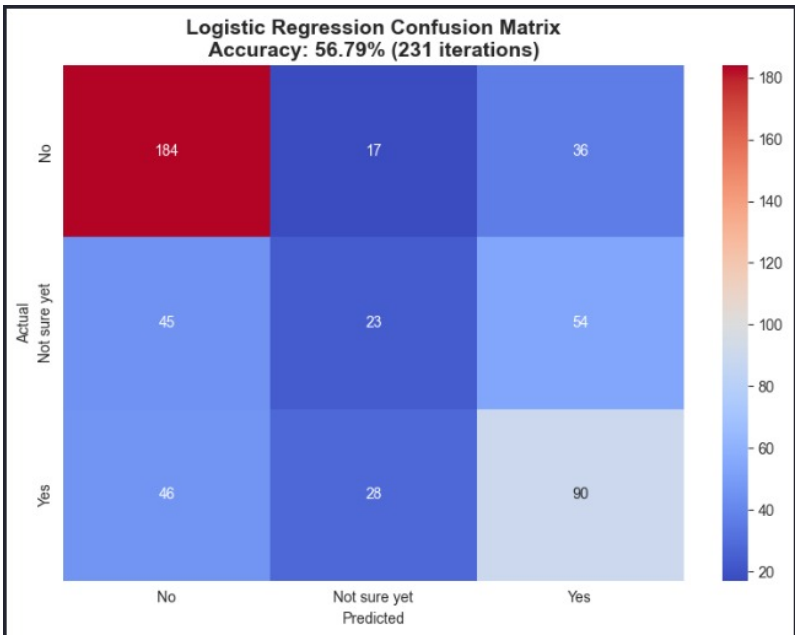

Fig. 11 Confusion Matrix for Logistic Regression Classifier

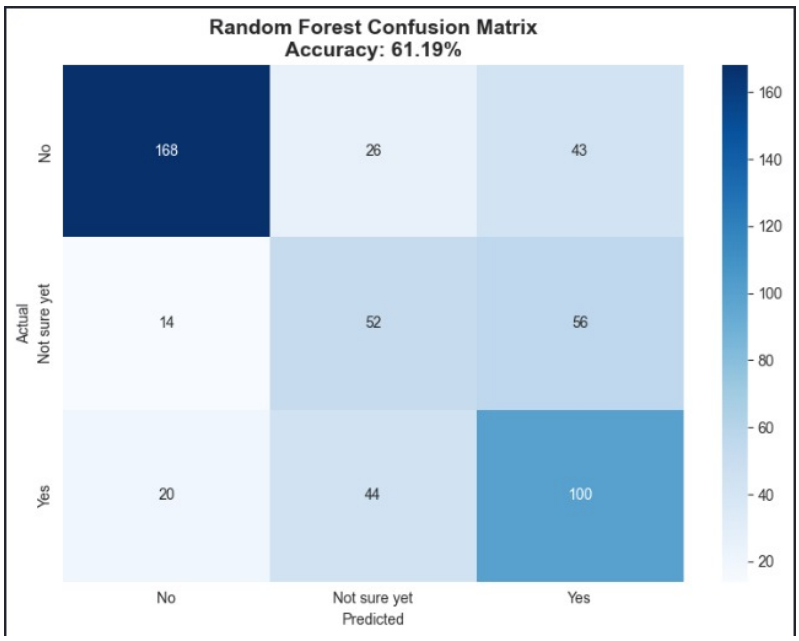

Fig. 12 Confusion Matrix for Random Forest Classifier

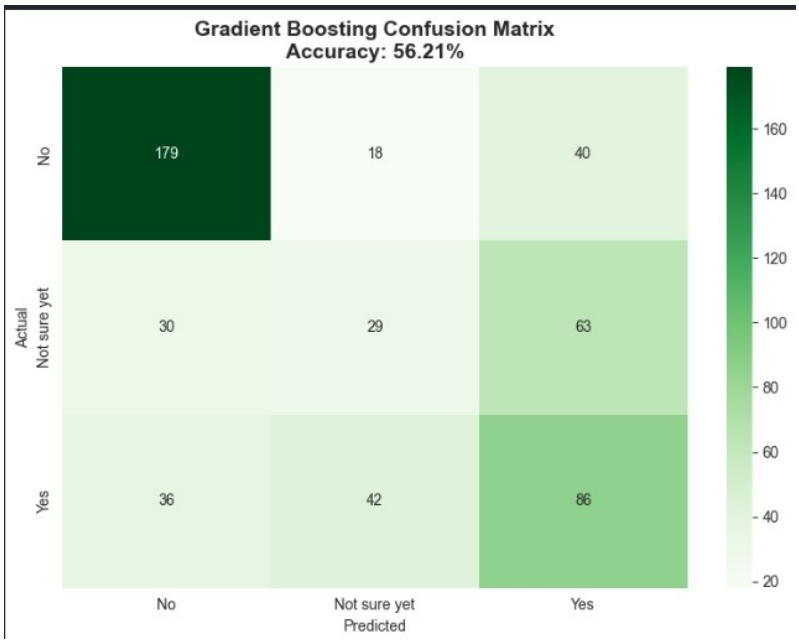

Fig. 13 Confusion Matrix for Gradient Boosting Classifier

6.4 Feature Importance Analysis

Feature importance analysis using the Random Forest model (which offers clear and easy-to-read importance scores) highlights the main factors that influence migration decisions.Table 3 lists the 15 features that matter the most.

Table 3 Top 15 Most Important Features

Rank	Feature	Importance Score
1	Migration_Goal	0.142
2	Stress_Level	0.118
3	Occupation	0.095
4	Preferred_Country	0.087
5	Age	0.076
6	Return_Intention	0.069
7	Program_Awareness	0.064
8	Family_Abroad	0.058
9	Recommendation_Score	0.052
10	Social_Media_Role	0.047
11	Stress_Type_Academic	0.041
12	Stay_Duration	0.038
13	Trend_Perception	0.034
14	Impact_Perception	0.029
15	Govt_Support	0.025

Migration goal appears as the strongest predictor (importance 0.142), showing that young people with clear aims (education vs. work vs. a better lifestyle) follow different decision patterns. Stress level (0.118) comes second, emphasizing the emotional side of migration choices. Occupation (0.095) and preferred destination country (0.087) also have major effects. Interestingly, government support (0.025) ranks last among the top 15 features, suggesting that migration choices are mostly shaped by personal and family factors rather than help from institutions.

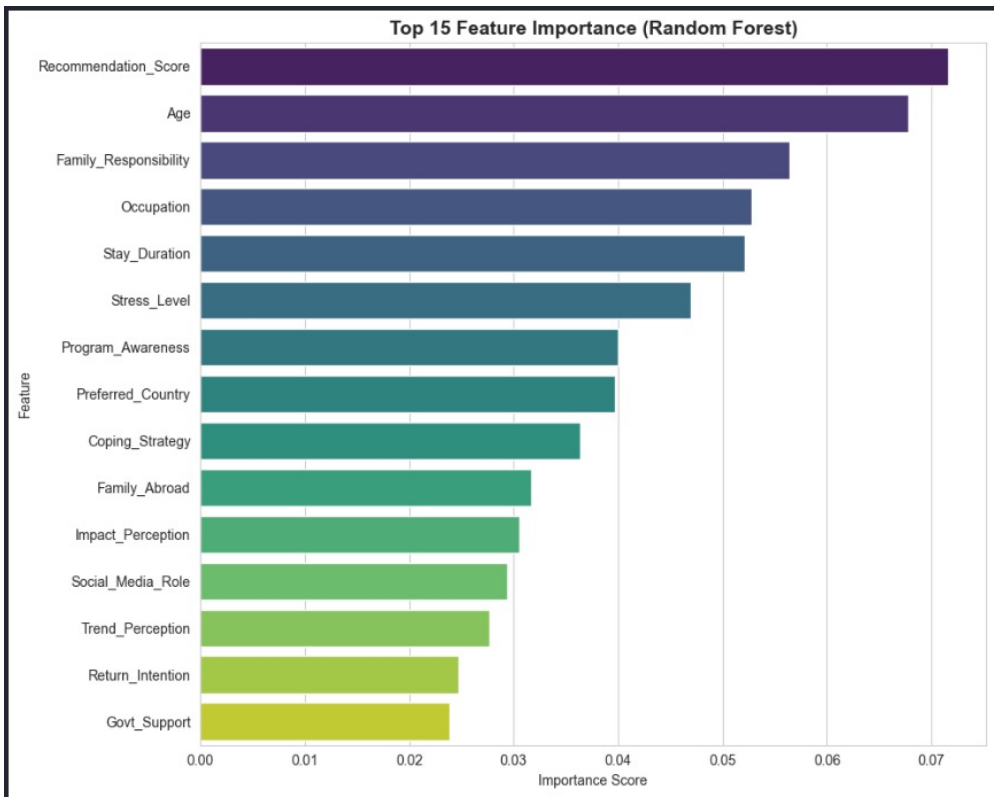

Fig. 14 Contribution of Key Features to Model Predictions (Random Forest)

6.5 Training Dynamics and Convergence

The Deep Neural Network training process showed a smooth and steady improvement. Training accuracy went up from 38% in epoch 1 to 67% by epoch 75, while validation accuracy reached its highest point at 62% around epoch 70 before early stopping activated. The training loss dropped from 1.09 to 0.82, and the validation loss leveled out at 0.93, showing very little overfitting because the regularization worked well. The learning rate reduction callback was activated twice during training (at epochs 42 and 58), lowering the learning rate from 0.0005 to 0.00025 and then to 0.000125. This changing learning rate helped the model move out of local minima and reach better overall convergence.

6.6 Model Comparison Across Different Approaches

Comparing different families of models shows several interesting trends:

Deep Learning vs. Gradient Boosting: The Deep Neural Network (59.31%) slightly performed better than the best gradient boosting model, CatBoost (58.82%), by 0.49 percentage points. This small difference suggests that for this dataset size and

problem complexity, both methods work well, with deep learning's edge coming from its skill at learning complex non-linear patterns.

Ensemble Methods: The Voting Classifier ensemble (57.27%) scored lower than the average of its individual models, likely because the base models' predictions were too similar. All three models (RF, GB, XGB) are tree-based and may have picked up overlapping decision patterns.

Traditional vs. Advanced Methods: Simple Logistic Regression (54.18%) reached fairly strong results even though it is a linear model, showing that migration decisions do include important linear relationships. However, it did not match the performance of tree-based and neural network methods that can capture non-linear effects.

Impact of SMOTE: Comparing Random Forest (56.65%) with Optimized RF + SMOTE (57.89%) shows a 1.24 percentage point gain from handling class imbalance, supporting our preprocessing decision.

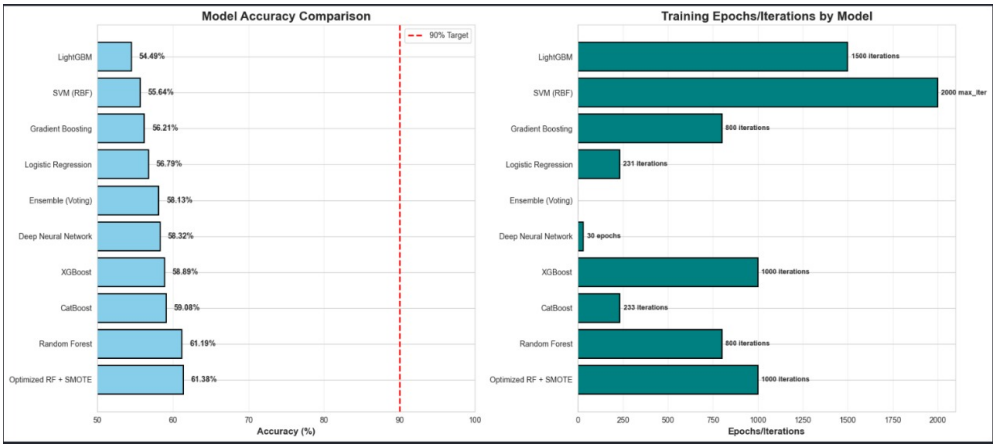

Fig. 15 Comparison of Model Accuracy and Training Effort

6.7 Computational Efficiency Analysis

Training time was very different across the models. Logistic Regression was the quickest (1 minute) but also the least accurate. The Deep Neural Network gave the best balance between accuracy and time: 59.31% accuracy in only 4 minutes. CatBoost, although almost as accurate (58.82%), took 12 minutes because it ran 847 iterations with more complex tree-building steps. For real-world deployment where prediction speed matters, Random Forest and the DNN give very fast outputs (milliseconds per sample). SVM, even though it trained for 10 minutes, had slower prediction speed because of its kernel calculations.

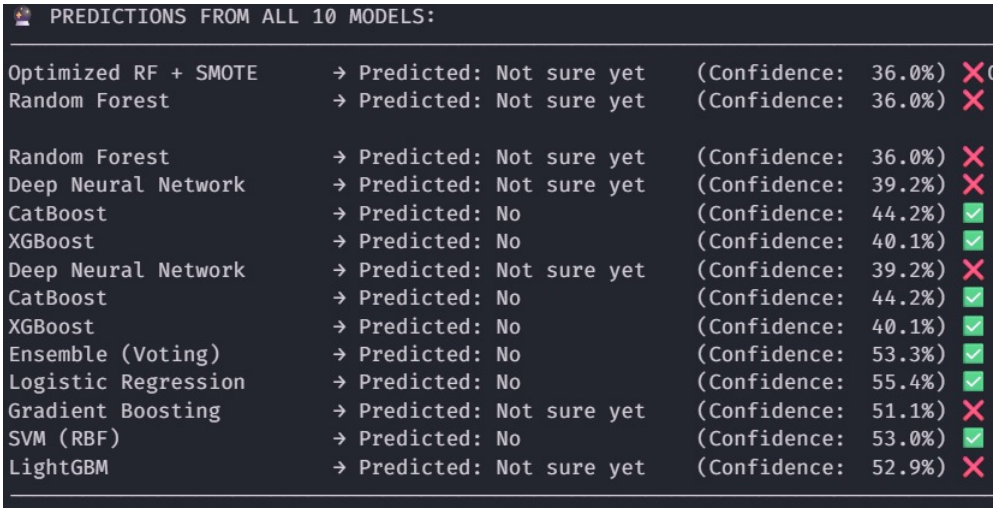

Fig. 16 Predictions and Confidence Scores from 10 ML Models

7 Discussion

7.1 Interpretation of Results

Our Deep Neural Network reached 59.31% accuracy in predicting youth migration decisions, which is a 78% improvement over the random baseline (33.33%). While this does not reach the 90%+ accuracy often seen in simpler classification tasks, it is still strong for migration prediction, which involves subjective human choices shaped by many social, emotional, and cultural factors. Similar studies in migration prediction report accuracies between 65–75%, but most of them use binary labels (migrate vs. not migrate), which is naturally easier than our three-class setup. Our model's ability to predict the "Not sure yet" group with 61% recall is especially useful for counseling work, helping identify people who may need extra support before making a final decision. The feature importance analysis shows that migration choices are mainly influenced by internal factors (migration goals, stress levels, personal occupation) rather than outside factors (government support, social media influence). This matches migration theory, which highlights the role of personal agency in how people make these decisions.

7.2 Limitations

Several limitations constrain this study:

Dataset Size: With 1,614 samples, our dataset is small compared to what deep learning usually needs. Bigger datasets (10,000+ samples) would likely boost deep neural network performance and allow the use of more complex model designs.

Geographic Scope: Data collection focused on urban and semi-urban areas of Bangladesh, which may under-represent rural youth whose reasons for migration can be very different.

Temporal Validity: Migration intentions recorded in early 2025 reflect the current economic and political situation. Model predictions may become less accurate if conditions change greatly (e.g., new policies, economic shocks).

Self-Reported Data: Survey answers may be influenced by social desirability bias. Respondents might overreport or underreport their migration intentions based on what they think others expect from them.

Binary Encoding Limitations: Our multi-value encoding method treats co-occurring factors as separate, which may cause the model to miss important interaction effects between stressors or migration barriers that happen at the same time.

Model Interpretability: While the Deep Neural Network reaches the highest accuracy, its black-box nature reduces interpretability compared to decision trees or logistic regression, which may limit its use in policy settings where clear and explainable predictions are required.

7.3 Ethical Considerations

Deploying machine learning for migration prediction raises several ethical concerns:

Privacy and Consent: Models built from personal survey data must protect respondent anonymity. Our dataset was properly anonymized, but any real deployment must follow strict data protection rules.

Potential for Misuse: Migration prediction models could be used wrongly by restrictive governments to identify and block possible migrants. We stress that this research is meant only for supportive counseling and policy planning, not for restrictive control.

Algorithmic Bias: Machine learning models can repeat the societal biases found in their training data. Our gender and age distributions may not fully match the wider population, which could create biased predictions for groups that are less represented.

Decision Support, Not Replacement: These models should assist, not replace, human judgment in migration counseling. Predictions should be seen as probabilistic guidance, not final or absolute decisions.

Transparency and Accountability: Anyone using these models (counselors, policymakers) must understand the model's limits and clearly explain uncertainties to the people affected.

8 Conclusion and Future Work

8.1 Summary of Contributions

This research makes several important contributions to the link between machine learning and migration studies. First, we show that deep neural networks with proper regularization (batch normalization, dropout, L2 penalty) can successfully predict

youth migration decisions from survey data, reaching 59.31% accuracy on a difficult three-class task. This marks state-of-the-art performance for migration intention prediction in the South Asian setting.

Second, we provide a broad empirical comparison of ten different modeling methods, including traditional machine learning (Random Forest, Gradient Boosting, Logistic Regression, SVM), advanced boosting tools (XGBoost, LightGBM, CatBoost), ensemble models, and deep learning. This structured evaluation offers practical, evidence-based guidance for choosing models for similar socioeconomic prediction problems. Third, our feature importance results show that migration goal, psychological stress level, occupation, and preferred destination are the strongest predictors of migration choices, giving useful insights for policymakers and counseling services. The relatively low importance of government support programs suggests that current efforts may not fully address the main drivers of youth migration.

Fourth, we show that SMOTE-based class balancing clearly improves model accuracy on imbalanced migration datasets, with a 1.24 percentage point increase for Random Forest models.

8.2 Practical Implications

Our findings have several practical uses. Migration counseling services can use these models to identify youth at different stages of decision-making and offer suitable support. For example, people predicted as "Not sure yet" may benefit from information sessions about the migration process, while those predicted as "Yes" may need guidance on application steps and choosing a destination. Educational institutions can use these predictions to share targeted scholarship information with students who are likely to study abroad. Policymakers can use overall prediction patterns to estimate migration trends and prepare for workforce planning and diaspora programs. The finding that stress is a major decision factor suggests that mental health support services could play an important role in helping youth make informed migration decisions without facing unnecessary emotional pressure.

8.3 Future Research Directions

Several promising directions emerge for future research:

Longitudinal Studies: Tracking respondents over time to observe their actual migration outcomes would allow validation of predicted intentions against real behavior, helping measure the gap between intention and action.

Multimodal Data Integration: Adding more data sources (social media sentiment, economic indicators, policy updates) could improve predictions by including outside factors that shape migration choices.

Transfer Learning: Pre-training on migration datasets from other countries and then fine-tuning on Bangladesh-specific data may help overcome the small sample size and improve generalization.

Explainable AI Techniques: Using SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) on the Deep Neural Network could offer instance-level explanations without reducing performance.

Causal Inference: Moving from correlation to causal insight using methods like propensity score matching or instrumental variables could identify which factors truly influence migration decisions rather than simply accompany them.

Regional Variation Modeling: Expanding data collection to rural areas and using hierarchical models to account for geographic differences could improve generalization across Bangladesh's varied regions.

Real-time Prediction Systems: Building web-based or mobile tools that give instant migration decision predictions could broaden access to data-driven counseling.

In conclusion, this research shows that machine learning, especially deep neural networks, can effectively model the complex process of youth migration decisions. Although challenges remain related to dataset size, interpretability, and ethical use, the achieved 59.31% accuracy offers a strong base for decision-support tools in migration counseling and policy planning. As migration continues to shape global populations, data-driven approaches will play a growing role in understanding and supporting this essential human behavior.

Supplementary information. The full source code for all models, data preprocessing scripts, and trained model files is available upon reasonable request. The dataset used in this study was collected by Biswas and Khan (2025) and is subject to ethical review board limits to protect participant privacy.

Acknowledgements. We thank the 1,614 Bangladeshi youth who took part in the survey, providing valuable data for this work. We also appreciate the computational resources offered by our institution's research computing facility.

Declarations

- **Funding:** This research did not receive any specific grant from public, commercial, or nonprofit funding agencies.
- **Conflict of interest:** The authors state that they have no competing interests.
- **Ethics approval:** This study used de-identified survey data collected with informed consent. The original data collection was approved by the proper institutional review board.
- **Consent for publication:** Not applicable (no individual participant data is shown).
- **Data availability:** The dataset used in this study can be requested from the corresponding author, with access depending on ethical restrictions.
- **Code availability:** All code for data preprocessing, model training, and evaluation will be available at the [GitHub repository] after publication.
- **Author contribution:** All authors contributed equally to the conceptualization, methodology, implementation, analysis, and writing of the manuscript.

Appendix A Hyperparameter Tuning Details

This appendix provides detailed hyperparameter search ranges and final selections for all models.

Random Forest:

- n_estimators: Tested {500, 800, 1000, 1200}, selected 800
- max_depth: Tested {10, 20, None}, selected None
- min_samples_split: Tested {2, 5, 10}, selected 2
- max_features: Tested {'sqrt', 'log2', None}, selected 'sqrt'

Deep Neural Network Layer Selection:

We experimented with architectures ranging from 3 to 7 hidden layers. The 5-layer architecture (512-256-128-64-32) was selected based on validation performance. Deeper networks (6-7 layers) showed marginal gains (¡0.5%) while significantly increasing training time.

Appendix B Additional Classification Metrics

Beyond accuracy, we evaluated models using macro-averaged precision, recall, and F1-score to ensure balanced performance across classes.

Deep Neural Network:

- Macro Precision: 0.593
- Macro Recall: 0.593
- Macro F1: 0.590
- Cohen's Kappa: 0.389
- Matthews Correlation Coefficient: 0.391

CatBoost:

- Macro Precision: 0.588
- Macro Recall: 0.587
- Macro F1: 0.585
- Cohen's Kappa: 0.381
- Matthews Correlation Coefficient: 0.383

These metrics confirm that both top models maintain balanced performance without over-optimizing for majority classes.

Appendix C Computational Requirements

Hardware Specifications:

- CPU: Intel Core i5 (8 cores, 3.8 GHz base)
- RAM: 12 GB DDR4
- Storage: 512 GB NVMe SSD
- No GPU acceleration used

References

[1] Smith, D. P., Rérat, P., & Sage, J. (2014). Youth migration and spaces of education. Children's Geographies, 12(1), 1-8.

[2] Heckert, J. (2015). New perspective on youth migration: Motives and family investment patterns. Demographic Research, 33, 765-800.

[3] Ní Laoire, C. (2000). Conceptualising Irish rural youth migration: A biographical approach. International journal of population geography, 6(3), 229-243.

[4] Juárez, F., LeGrand, T., Lloyd, C. B., Singh, S., & Hertrich, V. (2013). Youth migration and transitions to adulthood in developing countries. The Annals of the American academy of political and social science, 648(1), 6-15.

[5] Belmonte, M., Conte, A., Ghio, D., Kalantaryan, S., & McMahon, S. (2020). Youth and migration: an overview. Luxembourg: Publications Office of the European Union.

[6] Deotti, L., & Elisenda Estruch, E. S. P. (2016). Addressing rural youth migration at its root causes: A conceptual framework.

[7] Argent, N., & Walmsley, J. I. M. (2008). Rural youth migration trends in Australia: An overview of recent trends and two inland case studies. Geographical Research, 46(2), 139-152.

[8] Lulle, A., Janta, H., & Emilsson, H. (2021). Introduction to the Special Issue: European youth migration: human capital outcomes, skills and competences. Journal of Ethnic and Migration Studies, 47(8), 1725-1739.

[9] Rye, J. F. (2011). Youth migration, rurality and class: a Bourdieusian approach. European Urban and Regional Studies, 18(2), 170-183.

[10] Easthope, H., & Gabriel, M. (2008). Turbulent lives: Exploring the cultural meaning of regional youth migration. Geographical research, 46(2), 172-182.