National College of Ireland

Advanced Visa Outcome Predictions for Superior Accuracy and Interpretability

MSc Research Project
MSc Data Analytics

Kishore Nallasivam
Student ID: x23205962

School of Computing
National College of Ireland

Supervisor: William Clifford

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Kishore Nallasivam		
Student ID:	x23205962		
Programme:	MSc Data Analytics	**Year:**	2024
Module:	MSc Research Project		
Supervisor:	William Clifford		
Submission Due Date:	12/12/2024		
Project Title:	Advanced Visa Outcome Predictions for Superior Accuracy and Interpretability		
Word Count:	6636 **Page Count:** 20		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Kishore Nallasivam

Date: 12/12/2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Advanced Visa Outcome Predictions for Superior Accuracy and Interpretability

Kishore Nallasivam
X23205962

ABSTRACT:

The H1B visa plays an important role for skilled workers looking for employment in the U.S.; however, its application process is unclear and inconsistent, presenting challenges for employers. This study addresses these issues using two robust predictive models: the Bi-LSTM model for sequential data and XGBoost for structured data analysis to predict H1B visa outcomes with high accuracy and interpretability. This research applies advanced feature selection and data balancing methods to H1B visa data from the 2017 to 2022 fiscal years to address class imbalances and achieve highly generalized models. The deep learning and machine learning models are employed to find a factor influencing visa decisions. More complex sequential dependencies are generated with the help of Bi-LSTM, while enhanced scalability and interpretability are derived from XGBoost. As evaluation measures, accuracy, F1 score, and recall were adopted. These metrics show improved forecasting and efficiency, along with greater transparency in the decision-making process. It offers practical recommendations for applicants and immigration authorities while offering a starting point for applying predictive modeling to additional concept classes. To eliminate unfair practices in the issuance of visas, the study aims at making the process more transparent and efficient.

Keywords - Deep learning, H1B visa, Employment, visa outcome prediction, XGBoost

1. INTRODUCTION:

The US remains a most favorable country for many people who want to migrate there due to its array of visa options that are distinguished by the type of jobs. Visas such as L-1, E-3, EB-2, and O-1 meet different demands; yet the most popular visa is the H-1B (Wasem 2016). The H1B visa is considered the most popular because it enables people to work in skilled jobs that require at least a university degree (Chatterjee, Velpuru C T. Jagadeeswari 2021). The H1B visa remains highly competitive, with only 85,000 candidates selected each year under U.S. law (Pathak et al., 2023). This limit increases tension to applicants in a way that is no longer innocent, fair, or impartial but highly selective. In addition, applicants who apply for visas are worried about the inconsistency that marks visa issuing procedures. Previous works relied on more conventional forms of ML to make visa outcomes predictions, and those models were unable to deal with the growth in the size of the visa application dataset and the increasing nature of the problem in the past five years (Pathak, Rees-Jones, C. Sönmez 2022) Since the visa flows have made their processes more complex, such models have exhibited low accuracy and failed to determine all the drivers of visa decisions.

In this research, these challenges are focused on employing advanced models such as Bi-LSTM and XGBoost. Dombe et al. (2020) focus on efficiency of Bi-LSTM in handling sequential data, making it a right choice for predicting H1B visa outcomes, XGBoost is known for its efficiency in processing structured data and can complement deep learning models by providing faster and more scalable predictions. These models are more appropriate for dealing with some of the uncertainties of visa applications and therefore result in enhanced prediction. This

research also appears to be aimed not only at attaining very good levels of accuracy in estimating the approval decision for visas but also at uncovering the decision determinants of visas. Still, contrary to previous techniques, deep learning together with XGBoost's effective gradient boosting tools is predicted to provide better interpretability, favorable flexibility, and relatively improved capability to handle the intricacies of anticipated visa data. A gap that remains unaddressed in the current literature on visa predictions, particularly about reducing waiting times. This approach could extend its benefits not only to applicants but also to immigration authorities. Finally, the proposed methodology, which combines both the deep learning methods and the structured-data models such as XGBoost (Czarnowski et al. 2018), can be applied for the prediction of other visa categories that can be of value to the VFS for numerous visa prediction problems. These models give flexibility and the creativity that are required in increasing the transparency, reducing the time taken in processing visas, and increasing the approval period.

1.1 RESEARCH QUESTION

To what extent can Bi-LSTM and XGBoost reduce processing times for H1B visas while delivering more accurate approval predictions?

1.2 OBJECTIVE

To address the above research question, the following objectives were established and successfully pursued in this research created and assessed H1B visa outcome prediction models through neural-network-based Bi-LSTM with XGBoost max-margin learning techniques. The H1B visa data from Kaggle required advanced procedures for data cleaning along with preprocessing methods. The models' predictive power was measured through the implementation of confusion matrix together with accuracy, F1 score and recall metrics for optimal prediction accuracy. Analyzing influential visa decision factors to improve transparency and clarity throughout visa examination. Future work toward better predictions should concentrate on processing speed improvements as well as improved decision-making optimization although this study primarily focused on prediction accuracy and interpretability.

1.3 MOTIVATION AND PURPOSE

Applicants face several challenges when searching for jobs and then applying for H1B visas. This project seeks to help prospective applicants identify employers that are more likely to sponsor their visa applications based on a combination of factors, including the candidate's profile and various employer parameters. Entering the USA legally is a process considered by many critical visa factors, beyond just securing a job offer. In response to this, the H1B visa application data from 2017 to 2022 is crucial, and this paper examines the trends in H1B visa applications during that period. Many individuals consider that obtaining a visa to enter the USA is nearly impossible. However, understanding the factors that influence a visa sponsorship can simplify this process. The goal of this paper is to develop a predictive model that predicts an employer's chances of securing a visa, thereby recommending U.S. employers who may be willing to hire them legally. Besides, this paper aims to enhance understanding of the factors considered by immigration officials when approving or denying visas. It can enhance the understanding of mechanism and transparency at every stage of an application. Eventually, this project seeks to ensure that skilled workers find suitable employers in the USA. This research is outlined to make the visa application process more transparent and reduce difficulties for prospective candidates.

2. LITERATURE REVIEW H1-B VISA PREDICTION 2017-2022

In recent research, the analysis of H1B visa outcomes incorporates statistical analysis, traditional machine learning techniques, and advanced deep learning approaches. Scholars have applied a range of analytical tools, including regression analysis, to analyze aspects that determine visa outcomes. Prediction models like Decision Trees, K-Nearest Neighbors, Random Forest, and SVM have moderate accuracy but don't scale well with large datasets that have many features. Nevertheless, Bi-LSTM and other neural networks outperform other models in capturing dependencies and sequential features of the visa application data. Other techniques, such as ensemble methods, have continued to improve the predictive accuracy in data with structure and imbalance in data.

2.1 Trends and Challenges Associated with The H1b Visa Outcome Prediction

Applicants must adjust their approaches for getting an H1B visa, given the doubts and limitations of the selection process. Many challenges that were highlighted by Dombe, Rewale, & Swain (2020) highlight challenges such as visa limits, salary, work location, and employment status, all of which affect approval outcomes. All these factors influence the likelihood of success for applications but introduce variance, complicating the process for both applicants and employers, making the process challenging for both applicants and employers, which established that artificial neural networks (ANNs) effectively analyze complex variables, achieving higher levels of accuracy in predicting H1B visa outcomes. To these challenges (Raj, Piri, Eluri, et al. 2023)included the H1B procedural and financial requirements of applying for the visa. The extensive paperwork often creates confusion about an applicant's job and legal status, preventing many from applying. Also, high legal and processing costs are an issue that poses a significant challenge to the implementation of the act, especially to small employers, simply because they may not be able to afford to incur such costs. The study also points to the fact that in analyzing trends in the most recent years between 2020 and 2022, machine learning techniques can be applied to offer insights that may make the process of visa application successful for both the applicant as well as the employer.

(Raj, Piri, Reddy, et al. 2023)study analyzes the challenges faced by H1B applicants between 2015 and 2019, and it also points out the fact that other than the inability to find jobs and lack of legal protection, This approach helped to avoid using participants names and made it easier for the researchers to find patterns in the visa outcomes and use them to make better predictions for stakeholders while using data science techniques. Their work showed that through the application of data analysis, it was possible to minimize the amount of uncertainty that surrounded H1B petitions and at the same time lessen the costs accrued and stress experienced by applicants and employers.

Lastly, Chatterjee, Velpuru, & T Jagadeeswari (2021) explored the poor approval rates concerning this permit in the current world where the competition is high. Given that a significantly small number of applications has been approved because of increasing indispensability and stringent assessment requirements, applicants experience intensifying pressure for identifying the factors affecting approval outcomes. The study also addressed a technical issue of how to prepare diverse and intricate data to be usable on the machine learning algorithms. In this paper, the author has achieved a 94% accuracy rate in visa approval predictions, determining that advanced techniques can effectively analyze the H1B visa process. These models ensure that applicants can gain insights on how they may be able to overcome these barriers as well as update the process of application. The studies highlight that predicting H1B visa outcomes requires addressing procedural challenges leveraging neural network techniques.

2.2 Traditional Machine/Deep Learning Methods for Visa Predictions:

In the H-1B visa prediction, the traditional machine learning and deep learning models are applied, and each model has its advantages and disadvantages. Research by Prateek & Karun (2019) together with Thakur et al. (2018) evaluated traditional machine learning models such as Decision Tree, C5.0, Random Forest, Naive Bayes, Neural Networks and Support Vector Machine (SVM) for their applications in H1B visa predictions. Tests using the C5.0 model produced the most accurate results of 94.62% when compared to other predictive models. Despite their benefits these models still present issues that revolve around model fitting limitations and the need for adjustment tuning, which shows the effectiveness of the model in predictive tasks compared to boosting mechanisms. But issues like overfitting and very high reliance on the best choice of the parameter continue. This paper (Son, Byun & Lee 2018) focuses on a classifier of application metadata for the assessment of visa approval. In this classifier, simple and efficient operation is defined by such critical attributes as job descriptions, wages, and metadata from applications. Yet, this paper does not compare these methods to more standard machine learning models and leaves significant questions in terms of how these approaches work in large data sets. (Swain et al. 2018) turn a spotlight on ANNs as models with the ability to process the relationships and non-linear patterns within visa data. It means that ANN models can figure out the shift in certain parameters, like changes in economics and policies that affect trends in H-1B visas Deep learning models with their traditional 'black box' reputation have received increased interpretability from XAI developments during the past decade. Medical imaging benefits most from explainable AI methods described by Samek et al. (2017) and their saliency mapping methods help reveal how AI makes decisions in complex systems. According to Huff et al. (2021) XAI technology plays a key role in promoting both accountability and fairness when used for high-stakes applications including visa outcome predictions. Responsible decision systems for both lawmaking and regulatory frameworks require these advancements to deliver interpretable and equitable decisions. In addition, ANNs have one disadvantage in terms of needing large training data sets to prevent the network from underperforming, which is a problem where the number of data sets is limited.

Finally, Thakur et al. (2018) on benchmarking papers assess the traditional classification models, including Decision Tree, C5.0, Random Forest, Naive Bayes, Neural Networks, and SVM. C5.0 gave an accuracy of 94.62% and the overall highest accuracy of 95.40% by the model combining multiple classifiers by the 10-fold cross-validation. Readers will find from this study that that Improving the prediction model performance by combining the strengths of many models is achievable, but there arises a problem of computational complexity. However, the absence of investigating more elaborate deep learning techniques analyzes the current state research gap in visa prediction techniques.

In conclusion, as simple as it might seem, it is advisable to approach it with caution since, even though they give accurate results, they lack the ability to handle large data and cause overfitting. Deep learning models like ANN are useful for nonlinear and compact data, but they are costly and need high data sets to work on. Therefore, the combination of characteristics of these methods could yield the best solutions for the visa predictions. While the traditional machine learning models offer moderate accuracy, deep learning models like ANN and Bi-LSTM are better suited for complex and large datasets.

Strengths and Limitations

Model	Strengths	Limitations	Accuracy	Author
Decision Tree	* Easy to interpret. * Handles numerical and categorical data efficiently.	* Overfitting with complex datasets. * Limited feature interaction handling.	**88.45%**	Karun et al. (2019)
Random Forest	* Reduces overfitting by averaging multiple decision trees.	* Less interpretable than decision trees.	**92.34%**	Karun et al. (2019)
C5.0	* Boosting mechanism improves accuracy.	*Requires parameter optimization.	**94.62%**	Karun et al. (2019)
Naive Bayes	*Simple and fast for large datasets. *Effective with independent predictors.	* Poor performance with correlated features. Independence assumption often unrealistic.	**85.76%**	Karun et al. (2019)
Neural Network	*Captures complex, non-linear patterns. * Robust in high dimensional feature spaces.	* * Requires large data.	**90.23%**	Karun et al. (2019)
ANN	* Identifies intricate patterns and non-linear dependencies.	* Requires extensive training data. Less interpretable decision-making process.	-	Kushankur et al. (2018)
Metadata Classifier	* Focuses on application metadata for visa. *Simple to implement.	*Does not compare performance against other ML models. * *Limited scalability.	-	Byun et al. (2018)
Ensemble Model	* Combines multiple classifiers for improved accuracy. Better simplification and robustness.	* Computational overhead. *Limited exploration of deep learning methods.	**95.4%**	Pooja et al. (2018)

Figure 1 – Performance of Traditional Machine Learning Models in H1B Visa Predictions

2.3 Deep Learning and Ensemble Techniques of H1B Visa Outcome Accuracy

In the H-1B visa prediction, the traditional machine learning and deep learning models are applied. Nugroho et al. (2024), highlights how the combination of these elements amplifies predictive results and stability to achieve perfect markers on diabetes classification. An example of how SMOTE-Tomek can help in this case is the important problem of data imbalance directly relevant to growing numbers of underrepresented H1B visa case statuses, such as 'Denied' or 'Withdrawn.' Nonetheless, the difficulties in training multiple integrated models and the possible overfitting encourage more careful tuning and validation while using such an approach for visa predictions. Rane, Choudhary & Rane (2024) discuss stacking, boosting, bagging and will help to show how ensemble methods can help to handle the problem of imbalance and increase the level of prediction. These methods can be used as a supplement to Bi-LSTM and XGBoost since the former effectively processes sequential dependencies and the latter fits well for structured data. But problems such as overfitting and the necessity to engineer features remain. In utilizing highly sensitive applications such as the approval of visas, techniques like XAI should be used in order to make the decision-making process more understandable. In sentiment analysis, Kota & Munisamy (2022) show how the integration of CNNs for the feature extraction process and Bi-LSTM for the sequence process leads to high levels of prediction accuracy through the assistance of attention mechanisms and word embedding. All these improvements could be applied to H1B data for handling textual features for instances like job titles and employer names. However, challenges like computational complexity and overfitting are certain things because models developed must be optimized for a certain domain of use.

Lastly, Yan et al. (2023) compare the Bi-LSTM and the HS-Bi-LSTM to show that the latter is hierarchical when controlling delicate dependence relations. HI and HS-Bi-LSTM both are good for sequential and time-series data, and HS-Bi-LSTM might prove even better for the complex data of the H1B applications. But the right combination of computational speed and model complexity becomes significant and more important in large datasets.

Overall, it is seen that the methodologies explained in these studies provide a potential way for enhancing outcome prediction of H1B visas, and for applying these methods, there are limitations and barriers like computational cost, overfitting, and utilization of feature engineering approaches that have been highlighted. If these approaches are applied together, it could easily lead to a large increase in the accuracy of prediction while not compromising on the overall model stability or sequence processing significantly improves prediction accuracy, assisted by attention mechanisms and word embeddings Improvements such as attention mechanisms and word embeddings could enhance the processing of textual features like job titles and employer names in H-1B data. However, problems such as computational costs and overfitting reveal the importance of fine-tuning the models to the various domains of operation.

3. METHODOLOGY FOR US VISA OUTCOME PREDICTION

This study employs the CRISP-DM (Cross-Industry Standard Process for Data Mining) model to analyze H1B visa data from 2017 to 2022. This flexible approach is suitable for building and updating effective models, which is why it is used for integrating deep learning (Bi-LSTM) within the study and machine learning (XGBoost). The aim of these models in this study is to predict the outcomes of visa cases using these features of the case and the applicant, employer, wage, and job details. The next sub-section explains the overall approach taken to clean and transform the data and the way the various models were trained to predict visa outcomes.

The data comprised H1B visa applications collected from Kaggle for the years 2017 to 2022. The raw data was provided in CSV format, including variables such as Visa Type, Employer Name, and Visa Case Status. The combined total dataset includes 3,980,349 records where data from multiple years have been used to amass this number.

3.1 Data Preprocessing

Data preprocessing is essential for preparing the dataset so that algorithms can easily recognize it. The dataset was divided into six subsets corresponding to the years 2017, 2018, 2019, 2020, 2021, and 2022.

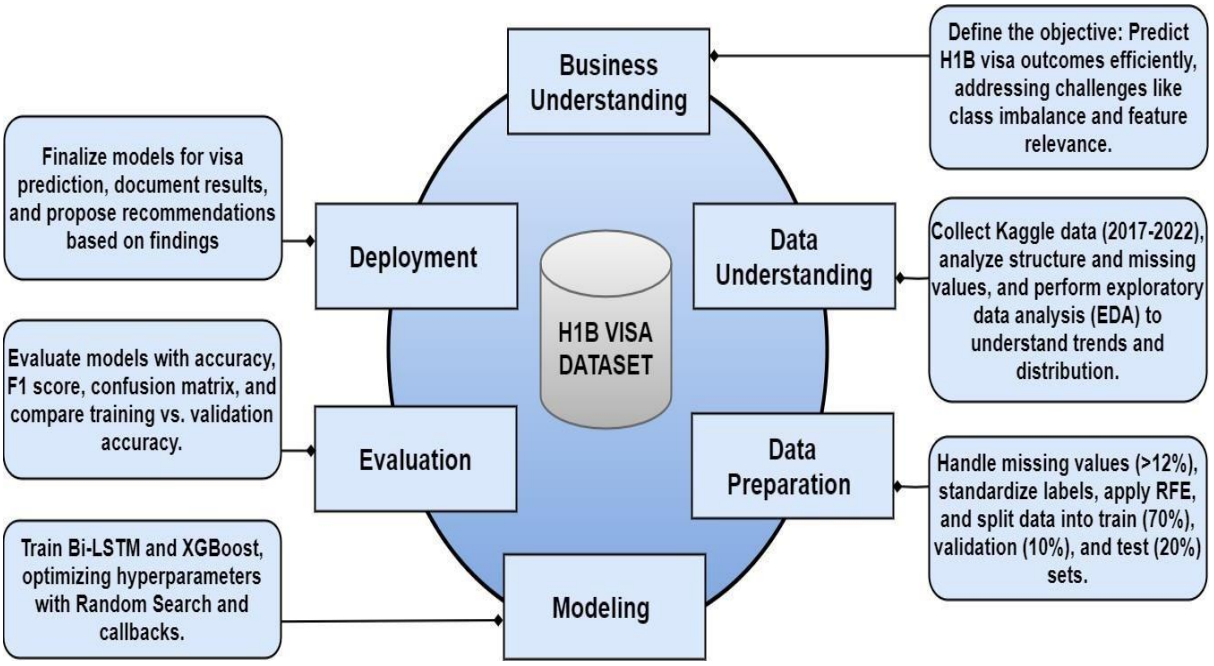

Figure 2 – Design process flow

Features	Descriptions
Visa Type	Specifies the visa category (e.g., H1B, L-1) applied for by the worker.
Employer Name	Name of the employer sponsoring the visa
Job SOC Title	Standard Occupational Classification title for the job role
Job Title	Job title designated by the employer for the position
Full Time Position (Y/N)	Indicates if the role is full-time (Y) or part-time (N)
Worksite Location	Location where the work will be performed
Prevailing Wage (USD)	Government-mandated minimum wage for the job location and role.
Unit of Pay	Pay frequency (e.g., hourly, annually) for the wage specified.

Employer Location	Primary address or headquarters of the employer.
Employer Country	Country in which the employer is based.
Visa Case Status	Current outcome of the visa application (e.g., Approved, Denied).
Application Year	Year in which the visa application was filed.
Wage_Rate_Of_Pay	Specifies the actual wage offered by the employer for the position.
Quarter	Indicates the quarter of the year in which the visa application was filed

Figure 3 – Features Explanation

To this end, a new column, Year, was included in each subset to capture the application year. The data was then merged with a final data set with 3,973,349 rows and 14 columns. The following preprocessing tasks were performed:

- Handling Missing Values: Any column with more than 12% missing values was removed to avoid greatly compromising the overall data quality and unfairly skewing the model results. This threshold was chosen because attributes such as wage rate of pay (83.27% missing) and quarter (48.92% missing) would negatively impact the model. We set this limit to strike a good balance as to the data size to retain and the authenticity of records that would help solve the case.
- Data Cleaning: Removed irrelevant columns (e.g., 'Quarter').
- Label Standardization: The Visa Status Case column was normalized into two other categories: Certified and These labels were then encoded for model use with 0 representing Denied and 1 representing Certified, and so on.
- Feature Selection: To facilitate the selection process, Recursive Feature Elimination (RFE) was used to choose the 12 most useful features, which were selected as vector features, including Prevailing Wage, Job SOC Title, Full-Time Position, and Worksite Location, due to their significant influence on visa outcomes.

The final dataset will be chosen for modeling and contains 38,460,026 instances with 10 features. To train, validate, and test the model more effectively, the data have been divided into training, validation, and test sets. This is done with the help of a function called train_test_split in the scikit-learn library, so the target variable y maintains the distribution of classes across the splits.

- Training Set: 70% of the data, used to train the model.
- Validation Set: 15% of the data (50% of the temporary 30% set), used for model tuning and selection.
- Test Set: 15% of the data (the remaining 50% of the temporary 30% set), used for evaluating the model's final performance.

3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the most important part, as it will involve finding out the distribution of the data set. It assists in the understanding of patterns, variations, and maybe trends. Such prior knowledge aids in analysis later. The data pertaining to the H1B visa application was analyzed through visual graphics and descriptive statistics to get insights from the information.
When looking at the distribution of the visa case status, we can see that 95% of all identified applications are certified, with more than 3.7 million approved applications. In the Denied status, there were almost 1,25,485 cases. suggesting that H-1B visa applications are not frequently rejected by the American government. This imbalance further supports the high approval status of H1B visa applications; most of the applicants receive certification. The significantly lower number of denied cases gives the impression that most of the applicants receive positive results from the process.

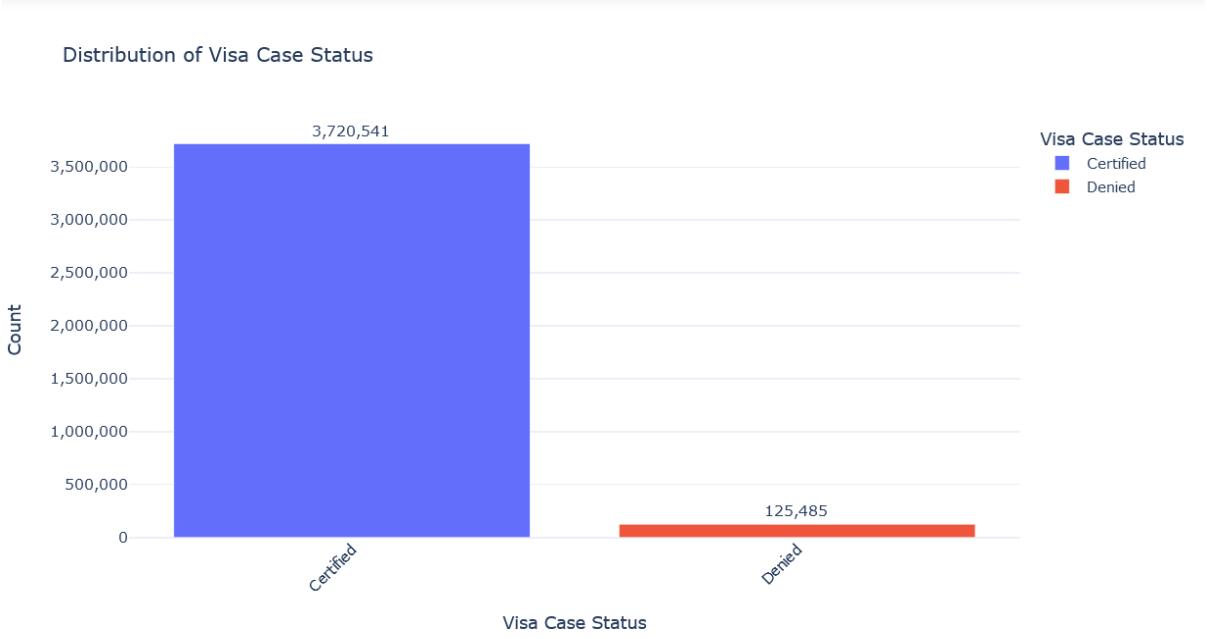

Figure 4: Distribution of the visa case status

The Figure 5 (yearly trends of H1B Visa 2017-2022) shows that the Certified cases constantly hold the highest number of applications, increasing in 2021 at over 800,000 before going down in 2022. In 2021, elevation may be due to an increase in demand or alterations in the standards of processing. The Certified – Withdrawn cases remain constant, which means there is not a great number of applicants that withdraw after passing initial process, most probably due to personal or employment situations. Both Withdrawn and Denied statuses show minimal variation and relatively low counts, with Denied cases indicating that outright rejections are rare. Overall, figure 5 suggests a high success rate for applicants meeting the eligibility criteria, with an outstanding increase in 2021 followed by a decrease in 2022, possibly due to external factors affecting visa demand.

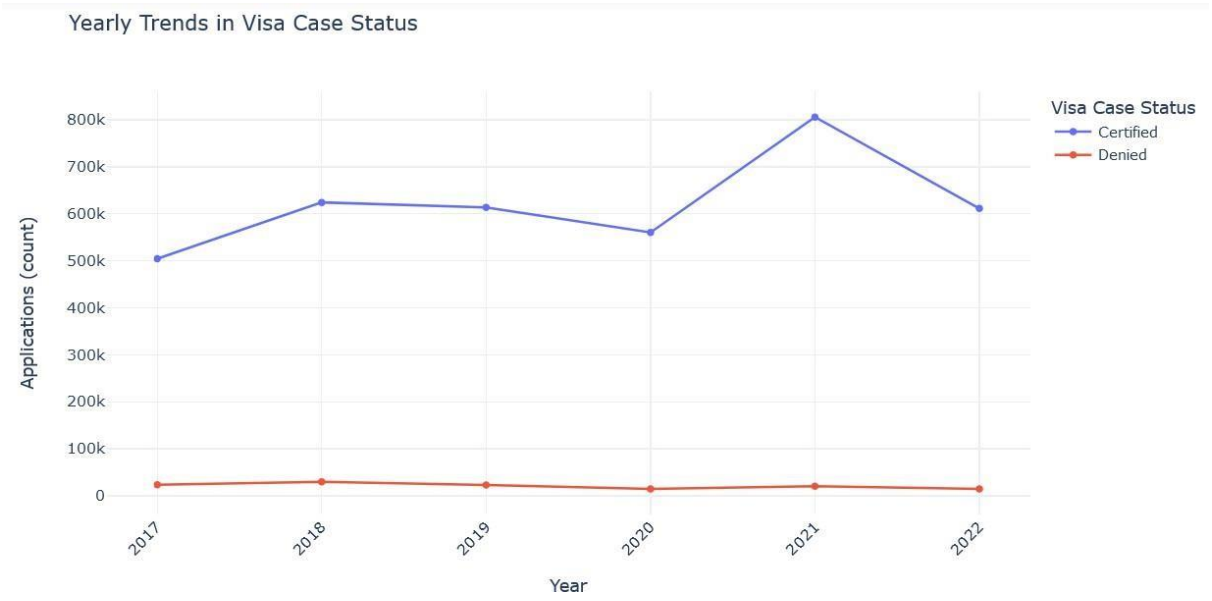

Figure 5: Yearly Trends in Visa Case Status

The information characterizes the case status of H1B visas for the years 2017-2022. Here the "Certified" status stays overall on top of the trend, while the number of certifications oscillates in the region of 500,000 to 600,000 during the years. Nonetheless, there seems to be a significant increase in the year 2021, with the application figure standing at 790,000. The distribution of the "Denied" category also seems quite balanced because it does not significantly change from one year to another. There are far fewer denied applications than certifications, and the outcome has hovered around 5,000-10,000 rejected applications each year. This has led to a general rise in approvals, more so in 2021; it could be due to an increase in demand or there was a change of policy.

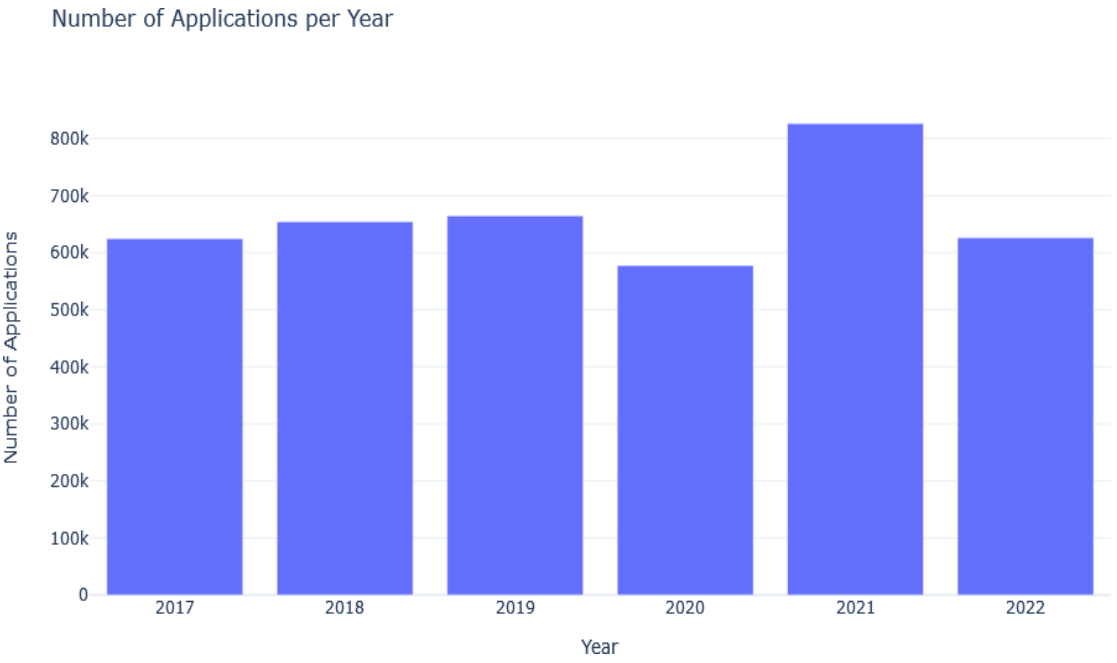

Figure 6: Number of Applications per Year

The distribution of H-1B visa cases from 2017 to 2022. Overall, the highest concentration of all cases is identified under the 'Certified' category, and there is a large fluctuation of around 800,000 applicants who were certified in 2021 and around 600,000 certified in 2022. Like "Denied" data is still a minority, but it vibrates slightly, indicating generally low values all the years. This trend could therefore signify better quality applicants or enhanced efficiency in the selection process, thus reducing applicant rejection. The fact that most of the certifications have been on the increase each year goes to show that the H1B visa approval system is very sound.

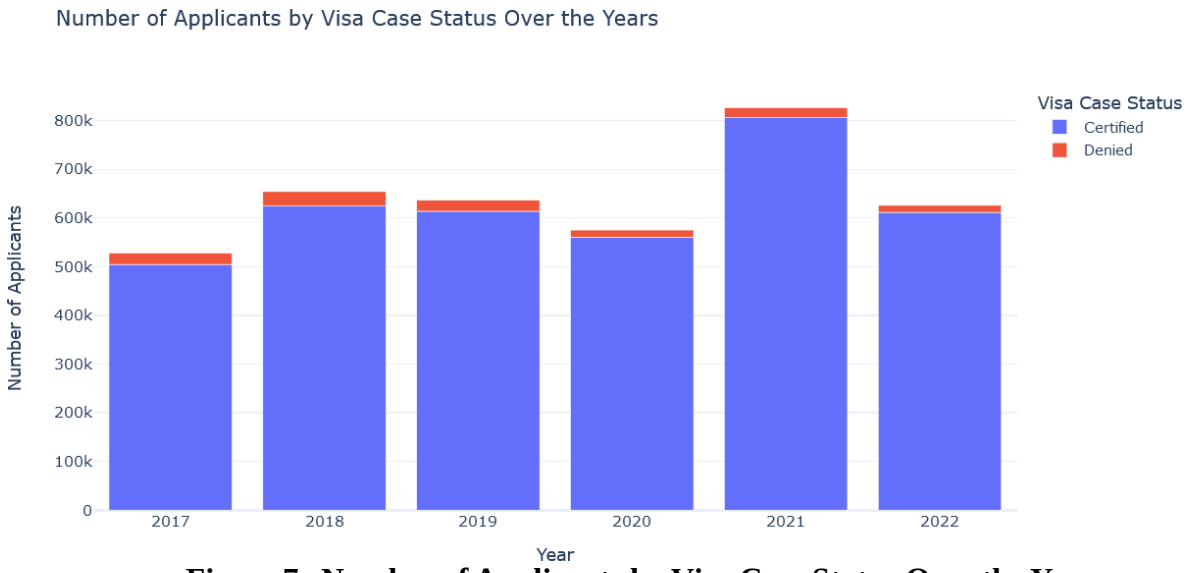

Figure 7: Number of Applicants by Visa Case Status Over the Years

3.3 Design Specification:

The project follows a two-layer methodology for analyzing visa outcomes. The layers have been updated as follows. The data preprocessing step was conducted using Jupyter Notebook. Data balancing, feature scaling and data transformation were used out of different preprocessing techniques. The pre-processed data was evaluated on both deep learning and machine learning, analyzing the compatibility and comparability between the models.

Bi-LSTM (Bidirectional Long Short-Term Memory):

This Bi-LSTM model is useful to handle sequential dependency in the datas. The architecture consists of:

Input Layer: The pre-processed data was reshaped into 3D tensor for sequential processing.

Two BiLSTM Layers:

- First BiLSTM layer with hyperparameter-tuned units (32–128) and a dropout layer (20%–50%).
- Second BiLSTM layer with a minor range of units (16–64) and a dropout layer (20%–50%).
- Dense Layer**:** Fully connected layer having an activation function of ReLU and then comes dropout in regularization.
- Output Layer: A sigmoid-activated neuron that is used to help predict binary visa outcomes, certified or denied.

Training and Hyperparameter Tuning:

- Hyperparameters, including the number of units, the dropout rate, and the learning rate, were tuned through Keras Tuner. The Adam optimizer was employed, with learning rates from [1e-3, 1e-4, 1e-5] being evaluated. Early stopping prevented overtraining in both the tuning and training phases.

3.4 XGBoost (Extreme Gradient Boosting):

To predict visa outcomes, a gradient-boosting decision tree was adopted for the classification of structured data.

Key Features of Implementation:

- n_estimators set to 100 for iterative
- max_depth restricted to 5 for controlling
- learning rate set at 0.1 to balance step size and
- subsample and colsample_bytree values were both set at 8 for optimal sampling.
- Log loss is used for the evaluation

4. Implementation, Evaluation, and Results of Bi-LSTM Models for US Visa Outcome Decision

This paper investigates the likelihood of US visa decision outcomes using the proposed Bi-LSTM model. The Bi-LSTM model is employed for binary classification, analyzing a dataset containing 11 features. Training was conducted on 3,178,679 samples, while testing utilized 794,670 samples. The target variable, visa case status, was grouped into two categories: certified and denied, following data preprocessing and balancing techniques.

Scaling:

StandardScaler was used to normalize the feature values, as there was a need to make sure that it had zero mean and unit variance. Large-scale ranges in some of the features could have caused the features to dominate the learning process; hence, the use of normalized data helped to remove this barrier in model performance.

Train-Test Split:

The dataset was then divided into train, validation, and test splits of 70%: 15%: 15% using the stratified split. This made it possible to have a check that all or any of the classes would have an equal number of data subsets as the other when you are dealing with sets that are imbalanced. The dataset presented itself as highly imbalanced; many cases were categorized as “certified.” To address this, the RandomUnderSampler, which is available in the imblearn library, was used.

The training set was balanced with 87,839 in each class (Certified and Denied). Similarly, the validation set was balanced to have 18,823 samples per class, ensuring an average evaluation across both classes.

Reshaping for Bi-LSTM:

Data was reshaped into a 3D tensor format: (samples, timesteps, features), where each sample had 1 timestep and 11 features. This format enabled the Bi-LSTM model to process sequential dependencies.

Model Architecture:

- Bi-LSTM Layers: Two bidirectional LSTM layers were added to capture long-range dependencies in both forward and backward directions:
- The first Bi-LSTM layer has a tunable number of units (lstm_units1) with a dropout rate (dropout_rate1).
- The second Bi-LSTM layer also has tunable units (lstm_units2) and dropout (dropout_rate2).
- Dense Layer: A fully connected layer with tunable units (dense_units) and a dropout rate (dropout_rate3) for feature extraction.
- Output Layer: A single dense unit with a sigmoid activation function for binary output.

Hyperparameter Tuning:

- We used Keras Tuner with a RandomSearch strategy to optimize the following hyperparameters:
- Number of units in the first and second Bi-LSTM layers (lstm_units1, lstm_units2).
- Number of units in the dense layer coined as Learning rate for Adam
- In total, 8 trials were run with 15 epochs each, though early stopping with validation

Best Model Training:

After tuning, the best hyperparameters were:

```
Best Hyperparameters: {'lstm_units1': 128, 'dropout_rate1': 0.2, 'lstm_units2': 32, 'dropout_rate2': 0.2, 'dense_units': 64, 'd
ropout_rate3': 0.2, 'learning_rate': 0.001}
```

- These hyperparameters, the model was trained for 20 epochs and early stopping of 3 epochs of no drop in the validation loss.
- The batch size was set to 64.

Optimization And Metrics:

- The model was compiled with the Adam optimizer, binary cross-entropy loss, and accuracy as the primary metric.
- Training history was recorded, including training and validation accuracy and loss across epochs.

Evaluation:

The Bi-LSTM model was evaluated using the test set, with key metrics including accuracy, recall, F1-score, and a confusion matrix.

Validation Performance:

- Validation accuracy during tuning: 66.65%.
- The learning curves indicate overfitting after the second epoch, as validation loss increased while training accuracy improved.

Test Performance:

- Accuracy: 65.79%.
- Weighted F1 Score: 0.767.
- Recall (for minority class): 50.74%.

The confusion matrix shows that it is difficult to achieve appropriate containment of false positive results and excessive sensitivity to false negatives.

```
Confusion Matrix:
 [[369986 188095]
 [  9272   9551]]
```

Learning Curves:

The plot of the training and validation accuracy show that the model had a high risk of overfitting after the second epoch. This is further supported by the form of the validation loss, for which the curve tends to rise sharply from the third epoch.

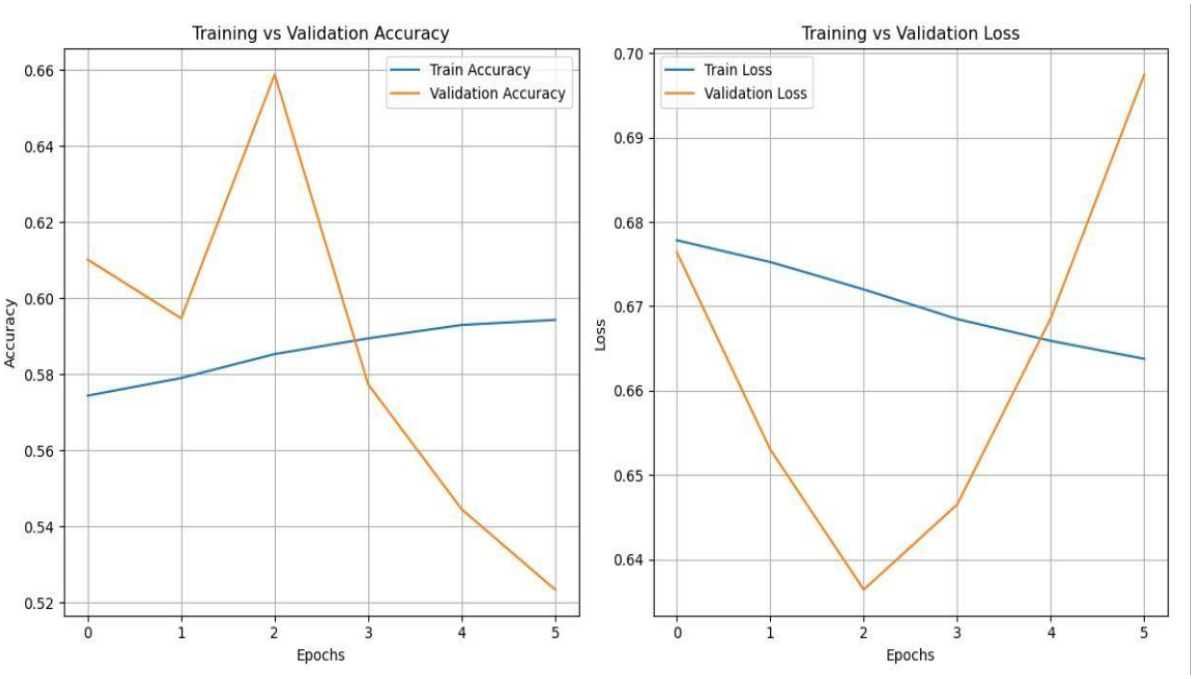

Figure – 8 Bi-LSTM model learning curves

4.1 Implementation, Evaluation and Results of XGBOOST Models for US Visas Outcome Decision

In this section, the likelihood of the US visa decision results is evaluated using XGBoost model in the decision-making process as it is believed to be one of the best models used in big data. XGBoost modeling is used in classification problem.

XGBOOST MODEL CONFIGURATION

The following hyperparameters were used for training:

- objective='binary:analysis: The model does a binary logistic analysis.
- n_estimators=100: Defines the number of boosting iterations possible.
- max_depth=5: It is common to set the maximal depth of trees in order to avoid getting overfitting.
- learning_rate=0.1: It fixes the step size shrinkage to prevent large updates.
- subsample=0.8: Choses 80% of the training data by randomly sampling from the whole training data for each boosting round.

- colsample_bytree=0.8: No of features used: 100 Features used for tree construction: 80.
- eval_metric='logloss': Calculates the loss on training and on the validation set during the training time.

MODEL TRAINING

- The model was trained on the balanced training set and on the validation set during the training time.
- Early stopping was used using validation loss in order to avoid overfitting.
- Training and validation log loss values were saved at every epoch

EVALUATION AND RESULTS

1. **Training and Validation Performance**

Both log loss for training and validation reduced from epoch to epoch, demonstrating effective learning over epochs.

- Training log loss: 0.63667 at the final epoch.
- Validation log loss: 0.64276 at the final epoch.
- Training vs validation accuracy curves show good convergence, with no evidence of overfitting.

2. **Test Set Performance**

The XGBoost model was evaluated on the test set, yielding the following results:

Accuracy: 60.39%

Confusion Matrix:

```
Confusion Matrix:
 [[336381 221700]
 [  6819  12004]]
```

The model performs well in identifying certified employees (class 0), it struggles with the minority class (rejected employees, class 1), resulting in higher false negatives.

5. ANALYSIS OF RESULTS IN THE CONTEXT OF THE DATASET

1. **Dataset Size and Class Imbalance:**

- The dataset comprises of 39,73,349 samples where most of them are certified (96.74%) while only a few are denied (3.26%), which was very difficult to handle while training the model, especially trying to predict the minority class, the ‘denied’. This could have made the model perform better for the majority class hence the relatively low recall score for rejected employees.

- This imbalance likely caused the model to favor the majority class, leading to a relatively low recall score for rejected employees.

2. **Comparison to Literature:**

- As highlighted in the literature review, some studies have reported achieving 94% accuracy using a different dataset with a more balanced class distribution and different feature sets.
- The observed performance gap in the present study can be explained by the high complexity, class imbalance, and large sample size of the dataset.

3. **Learning Curves:**

- Training accuracy improved throughout the epoch and validation accuracy started to decline slightly after the second epoch implying overfitting.
- Validation loss had a similar tendency, which strengthened this conclusion.

4. **Challenges and Future Directions:**

Compared to accuracy, the model struggled to identify the minority class as shown by the recall of only 50.74%.

Potential improvements include:

- Oversampling the minority class, for instance using SMOTE.
- Class-weighted loss functions to ensure that the minority class has high cost for misclassification.
- Specialized architecture is appropriate for cases where class imbalance is severe.

DISCUSSION

The features extracted from the Bi-LSTM and the performance of the XGBoost models give an understanding about the abilities and drawbacks of the choice of models to solve the binary classification problem of the study. The Bi-LSTM model utilized the benefits of its ability to learn sequentially and proved that in terms of temporal patterns in the data, it could yield good results. Nevertheless, it was less capable of training on unseen data probably due to poor management of overfitting and skewed classes. On the other hand, the XGBoost model shows reliable performance on structural data especially on selection of features and model interpretability. Nevertheless, the presented XGBoost model had certain issues in dealing with the minority class, thus making it difficult to deliver a balance of predictions.

The reliability of the outcomes is further enhanced by validation methods approved like log loss monitoring, and an assessment of the confusion matrix. However, the generality of these results is quite restricted owing to the intrinsic characteristics of the datasets and the preprocessing steps used in experiments. The generality of the results may be somewhat limited by the fact that the models depend on the selected hyperparameters and feature space.

The findings of this study suggest that the choice of model and preprocessing steps must be informed by the features of the data set Even though Bi-LSTM is effective with sequential data, XGBoost provides a more rigid structure in feature-based data. The future work could be done on the extension of both models in the use form where they both can be more beneficial when

used in conjunction with each other. Furthermore, additional efforts devoted to experimenting with other strategies of dealing with class imbalance, for example, cost-sensitive learning, and using ensembles, can improve the overall model performance. These are discoveries that will encourage more examination of viable frequencies of machine learning schemes appropriate for intricate, distorted data.

6. CONCLUSION AND FUTURE WORKS

In this study, other sophisticated models such as Bi-LSTM and XGBoost were used to classify H1B visa outcomes more accurately and interpretable. Nevertheless, class imbalance and dataset size were still prominent problems in the experiments with Bi-LSTM having a weighted F1 score of 0.767, and XGBoost model efficiency was significantly impressive with an accuracy of 60.39%. The results showed the problem with predicting the minority class while also pointing out that methods such as SMOTE or class-weighted loss function are necessary. These findings enhance the knowledge of nature and characteristics of the visa application processes and thereby provide a foundation for further research into how these models could be calibrated to enhance effectiveness and equity of decision-making. Future works can employ this approach into other visa categories or domains.

In future studies, it would be possible to define new and more effective interventions to tackle the drawbacks revealed in the current models, including class imbalance. With the help of approaches such as cost-sensitive learning or more sophisticated techniques of data augmentation, for instance, SMOTE, the performance of required predictions for the minority classes might be improved. In Bi-LSTM model, some works could try to include additional domain-specific embeddings or the use of more elaborated architecture such as attention mechanism may give better results. These can be modified on an XGBoost model like feature engineering where some of the features can be removed and hyperparameter tuning. Moreover, the integration of sequential learning ability of Bi-LSTM with feature-based capability of XGBoost can open trials for hybrid modelling.

REFERENCES:

1. Wasem, R.E., 2016. Temporary professional, managerial, and skilled foreign workers: Policy and trends.
2. Chatterjee, P., Velpuru, M.S. and Jagadeeswari, T., 2021. Success of h1-b visa using ann. In *Machine Learning and Information Processing: Proceedings of ICMLIP 2020* (pp. 491-499). Springer Singapore.
3. Pathak, P.A., Rees-Jones, A. and Sönmez, T., 2022. Immigration lottery design: Engineered and coincidental consequences of H-1B reforms. *Review of Economics and Statistics*, pp.1-43.
4. Czarnowski, I., Jędrzejowicz, P., Chao, K.M. and Yildirim, T., 2018. Overcoming "Big Data" Barriers in Machine Learning Techniques for the Real-Life Applications. *Complexity*, *2018*.
5. Anay, Dombe., Rahul, Rewale., Debabrata, Swain. (2020). A Deep Learning-Based Approach for Predicting the Outcome of H-1B Visa Application. 193-202. Available from: 10.1007/978-981-15-1884-3_18

6. Raj, P.B.A.S., Piri, J., Reddy, S. and Eluri, S.B., 2023, May. An Analytical Study of Regression Techniques towards H-1B Visa Prediction. In 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 871-876). IEEE.
7. Raj, P.B.A.S., Piri, J., Eluri, S.B. and Reddy, S., 2023, February. Work Visa Analysis using Machine Learning Techniques. In 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS) (pp. 616-621). IEEE.
8. Priyadarshini, Chatterjee., Muni, Sekhar, Velpuru., T., Jagadeeswari. (2021). Success of H1-B VISA Using ANN. 491-499. Available from: 10.1007/978-981-33-4859-2_48
9. Samek, W., 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.
10. Huff, D.T., Weisman, A.J. and Jeraj, R., 2021. Interpretation and visualization techniques for deep learning models in medical imaging. *Physics in Medicine & Biology*, *66*(4), p.04TR01
11. Prateek., Shweta, Karun. (2019). Predicting the Outcome of H-1B Visa Eligibility. 355-364. Available from: 10.1007/978-981-13-6861-5_31
12. Kyoungho, Son., Yung-Cheol, Byun., Sang-Joon, Lee. (2018). Prediction of Visitors using Machine Learning. Available from: 10.1109/ICIIBMS.2018.8549960
13. Debabrata, Swain., Kushankur, Chakraborty., Anay, Dombe., Ashitosh, Ashture., Nandakishor, D., Valakunde. (2018). Prediction of H1B Visa Using Machine Learning Algorithms. 2018, 1-7. Available from: 10.1109/ICACAT.2018.8933628
14. Pooja, Thakur., Mandeep, Singh., Harpreet, Singh., Prashant, Singh, Rana. (2018). An allotment of H1B work visa in USA using machine learning. International journal of engineering and technology, 7, 93-. Available from: 10.14419/IJET.V7I2.27.12642
15. De, Rosal, Ignatius, Moses, Setiadi., Kristiawan, Nugroho., Ahmad, Rofiqul, Muslikh., Syahroni, Wahyu, Irianandа., Arnold, Adimabua, Ojugo. (2024). Integrating SMOTE-Tomek and Fusion Learning with XGBoost Meta-Learner for Robust Diabetes Recognition. 1(1), 23-38. Available from: 10.62411/faith.2024-11
16. Nitin, Rane., Saurabh, P., Choudhary., Jayesh, Rane. (2024). Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions. 1(2), 18-41. Available from: 10.48185/smhs.v1i2.1225
17. Venkateswara, Rao, Kota., Shyamala, Devi, Munisamy. (2021). High accuracy offering attention mechanisms based deep learning approach using CNN/bi-LSTM for sentiment analysis. International Journal of Intelligent Computing and Cybernetics, Available from: 10.1108/IJICC-06-2021-0109
18. Xiaohui, Yan. (2022). ET and ET0: Bi-LSTM and HS-Bi-LSTM. Available from: 10.5281/zenodo.6774567