Hindawi
Complexity
Volume 2022, Article ID 8306473, 10 pages
https://doi.org/10.1155/2022/8306473

Research Article

Predicting Young Imposter Syndrome Using Ensemble Learning

Md. Nafiul Alam Khan,[1] **M. Saef Ullah Miah**,[2] **Md. Shahjalal**,[3] **Talha Bin Sarwar**,[4] **and Md. Shahariar Rokon**[5]

[1]*Institute of Mathematical Sciences, Faculty of Science, University of Malaya (UM), Kuala Lumpur, Malaysia*
[2]*Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang (UMP), Pekan 26600, Malaysia*
[3]*Department of Public Health, North South University (NSU), Dhaka, Bangladesh*
[4]*Department of Computer Science, Faculty of Science and Technology, American International University Bangladesh (AIUB), Dhaka, Bangladesh*
[5]*Applied Statistics and Data Science, Department of Statistics, Jahangirnagar University (JU), Savar, Bangladesh*

Correspondence should be addressed to Md. Nafiul Alam Khan; nafiul.nipun95@gmail.com

Received 4 September 2021; Accepted 7 January 2022; Published 21 February 2022

Academic Editor: Lingzhong Guo

Background. Imposter syndrome (IS), associated with self-doubt and fear despite clear accomplishments and competencies, is frequently detected in medical students and has a negative impact on their well-being. This study aimed to predict the students' IS using the machine learning ensemble approach. *Methods*. This study was a cross-sectional design among medical students in Bangladesh. Data were collected from February to July 2020 through snowball sampling technique across medical colleges in Bangladesh. In this study, we employed three different machine learning techniques such as neural network, random forest, and ensemble learning to compare the accuracy of prediction of the IS. *Results*. In total, 500 students completed the questionnaire. We used the YIS scale to determine the presence of IS among medical students. The ensemble model has the highest accuracy of this predictive model, with 96.4%, while the individual accuracy of random forest and neural network is 93.5% and 96.3%, respectively. We used different performance matrices to compare the results of the models. Finally, we compared feature importance scores between neural network and random forest model. The top feature of the neural network model is Y7, and the top feature of the random forest model is Y2, which is second among the top features of the neural network model. *Conclusions*. Imposter syndrome is an emerging mental illness in Bangladesh and requires the immediate attention of researchers. For instance, in order to reduce the impact of IS, identifying key factors responsible for IS is an important step. Machine learning methods can be employed to identify the potential sources responsible for IS. Similarly, determining how each factor contributes to the IS condition among medical students could be a potential future direction.

1. Introduction

Imposter syndrome (IS) is defined by a sense of not belonging, of being out of place, and of believing that one's perceived competence and success are underserved by others. Typically, this is regarded as a personal issue that should be addressed by keeping a record of accomplishments to serve as a reminder of progress [1–3]. The IS, which arbitrates the relationship between perfectionism and anxiety and partially influences perfectionism and depression, was first cited by clinical psychologists Pauline Clance and Suzanne Imes in late 1978 [4].

According to a more recent systemic review published in 2020, the prevalence of IS in the general population ranged from 9% to 82% [5], whereas another study conducted in 2020 showed that it varied from 22% to 60% among physicians and from 33% to 40% among trainee physicians [6]. According to the current IS research in the United States, 57% of pharmacy students [7] and 15% of medical students have IS [8]. Indeed, IS is becoming a significant public health concern on a global and regional scale. For instance, the prevalence of IS among medical students has been found to be 30% in the US [9], 45.7% in Malaysia [10], and 47% in Pakistan [11].

Several studies have found a substantial link between IS and overall psychological discomfort [5, 7, 12] and demographic and personal characteristics, including age [5, 7, 8], gender [6, 8, 13], and academic year [6, 10, 11]. The evidence indicates that medical students experienced moderated-to-strong IS [14], which has psychological and academic consequences [9–11]. Moving into clinical studies can be particularly difficult and lead to poor confidence in the students [15, 16]. Furthermore, several studies have found that IS has a detrimental impact on medical students' physical and intellectual well-being [10, 14]. As a result, individuals may miss out on chances since they are unaware of their capabilities.

These days, machine learning approaches help us to solve a variety of problems using machine knowledge. In addition to many algorithms, some of them have been shown to be the best for various purposes. There are various applications of machine learning algorithms, including health and medical issues, marketing, weather forecasting, and many more. For predicting IS, different statistical analysis and machine learning techniques can be applied to data collected from medical college students. Different statistical and machine learning methods can be utilized to build a predictive model in solving the classification, regression, and categorization problems [17]. Ensemble learning model yields better prediction in many other fields including sentiment analysis. To the best of the authors' knowledge, there is no published study on imposter syndrome among Bangladeshi medical students. Hence, this study proposes a solution to predict the students' IS utilizing the ensemble learning approach that incorporates random forest (RF) and artificial neural network (ANN). By utilizing the features obtained from RF and ANN, we train and test the final ensemble model. This type of approach is not thoroughly explored in the current literature that we have found so far. The results validate the improved performance of our proposed technique employing the four commonly used performance metrics, namely accuracy, precision, recall, and F-1 score. Therefore, the major contributions of this study can be summarized as follows:

(i) Populating a real-life dataset from the Bangladeshi medical students for predicting young imposter syndrome

(ii) Utilizing ensemble learning to predict IS among students in a course or school year, forecast whether a student is likely to suffer from IS

(iii) Evaluating the ensemble model for predicting the IS employing accuracy, precision, recall, and F-1 score

2. Materials and Methods

In this section, we followed the steps mentioned below for our research. Data collection is imperative for any type of study. In the data collection section, we showed our employed method of data collection and reasoning. Then, we moved on to the data processing step, where different methods of data cleaning and data tabulation are employed for further processing. Finally, we used two different types of machine learning techniques, that is, ANN (artificial neural network), RF (random forest), to generate an ensemble model for predicting IS. Figure 1 represents the overview of the methodology that is employed in this study.

2.1. Data Collection

2.1.1. Sample Size Determination. To achieve the desired sample size in our study, we used the snowball sampling method. The snowball sampling method is of the nonprobability variety. We used this type of sampling method in particular due to the lack of a sampling frame and the nonavailability of desired data. Using Equation (1), [18] we obtained the minimum sample size required for the study.

$$n_0 = \frac{z^2 pq}{d^2}, \tag{1}$$

where n_0 is the required sample size, z is the standard normal variate, p is the estimated proportion of the population with the attribute in the question and $p + q = 1$, and d is the allowed maximum error in estimating a population proportion. We considered the degree of accuracy, where $d = 0.05$ and $p = 0.47$, as the approximate proportion of the YIS, which was derived from a study conducted in Pakistan [11]. After calculation, we reached a minimum sample size of 382. However, we included more than 382 samples to reach a more accurate conclusion in our research. Finally, we managed to collect 500 unique samples for our study and discarded any incomplete questionnaires.

2.1.2. Study Design and Settings. This study was conducted using the cross-sectional survey to assess whether a student had IS or not. Our research was conducted among medical college students who study in Dhaka city. We did not include samples from any medical college outside of Dhaka city. Data were collected from 1st February 2020 to 30th July 2020. In our research, we collected samples from both public and private medical colleges which are situated in Dhaka city. Throughout our survey, we used snowball sampling (a nonprobability sampling method) to achieve desired sample size. We had to use this type of sampling method because of the lack of the sampling frame and the nonavailability of desired data. Participants of the study were students of both public and private medical colleges ranging from 1st to 5th year of study. We collected the sample data using face-to-face interviews. Respondents were reassured that all the information collected in our survey would be kept strictly confidential and would not be used for anything other than research purposes. Written consent was taken from the respondents. Then, we briefly introduced the student to our study and presented the questionnaire. The response was collected immediately. Initially, we identified some potential IS cases among our participants. Then, we asked them to refer others for the study. Afterward, we contacted them via e-mail or mobile phone and collected data after a face-to-face interview. The collected data were compiled for further processing and analysis. After cleaning the raw data, we had 500 complete observations for our study. We discarded any incomplete questionnaire.

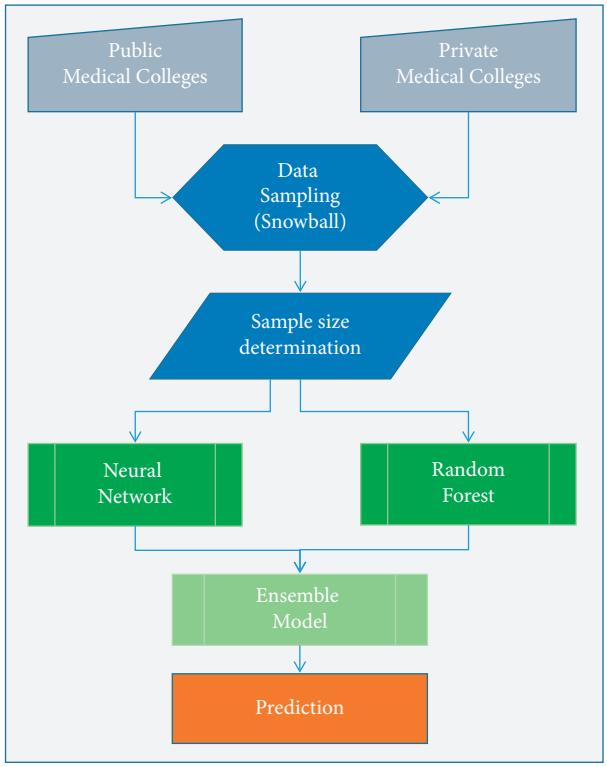

Figure 1: Employed methodology of young imposter syndrome prediction study.

2.1.3. Scale and Measures

(1) Dependent Variable. In our research, we considered having IS as the dependent variable. The dependent variable is categorized into two categories, that is, "Yes" (has IS) and "No" (does not have IS). "Yes" is denoted by 1, and "No" is denoted by 0.

The Young Imposter Syndrome (YIS) [9] scale was used to determine whether the person had IS or not. The scale was in the form of questions, and a student was considered to have imposter syndrome if they answered "Yes" to five or more of them. The YIS scale contains a total of eight items, and if a student scored on the YIS scale, then we considered him/her having IS otherwise, and he/she does not have IS. Detailed information on how respondents addressed all eight questions [9] is displayed in Table 1.

(2) Independent Variable. We considered respondents' sociodemographic and some specific academic data as independent variables. Gender (male, female), age group (18–21, 22–25), Body Mass Index (BMI), smoking status (current, past, never), and living with family (yes, no) were among the sociodemographic variables. The WHO BMI standard scale [19] was used to categorize BMI. The respondents' economic situation was determined by their monthly family income. Monthly family income (MFI) was divided into four categories: ≤ 20,000 BDT (approximately US$237), 21,000–30,000 BDT, 31,000–40,000 BDT, and ≥ 41,000 BDT. This MFI interval was a category that we created. The academic year was classified as first, second, third, fourth, or fifth, and the reasons for attending medical school were classified as personal preference, family preference, failure to qualify for the interesting department (departments in which they wanted to study except medical science), and better job facility (They chose to study medical science because of the better job facility it offers). Each response was divided into two categories (yes, no). All pertinent information was obtained directly from each participant.

2.2. Data Processing. After the interviewing process, we collected all questionnaires. Then, the collected data were compiled for processing and analysis by using statistical software (SPSS). First, we removed any duplicate or irrelevant data from our dataset for the sake of data cleaning. Then, we fixed any structural errors such as strange naming, typos, incorrect capitalization. Next, we validated the dataset according to the aim of the study. Finally, after cleaning and tabulation of raw data, we had 500 complete observations for our study purpose. We did not face any problem regarding missing value because we discarded any incomplete questionnaires from our study.

2.3. Predictive Model Generation. For developing the predictive model, the ensemble learning model [20] is utilized. The study utilizes weighted ensemble modelling using two machine learning approaches, namely random forest [21] and neural network [22]. With the same weight of 1, the final ensemble model is generated as the predictive model. The parameters and structures of all the models are provided below. For the predictive model's validation, K-fold cross-validation is employed instead of train test split as the train test split method introduces bias to the model, which is one of the reasons for the poor accuracy of any model. To eschew this problem, the five folds validation is employed in generating a predictive model.

2.4. Random Forest. Random forest (RF) is an ensemble learning method primarily based on ensemble-based decision trees [23, 24]. For selecting feature importance, OOB or out-of-bag data were used, and to evaluate feature importance, 2 steps were taken, that is, (1) two-thirds of the training dataset were used to build the predictive classifier, and the remaining dataset was used to evaluate the classifier's performance. (2) The decrease in prediction performance was used to assess the importance of each feature. We reported the performance evaluation in terms of accuracy and the Gini index. In RF, the Gini index is used to determine the ability of potential discrimination among features. The Gini index is defined as per Equation 2.

$$\text{Gini}_{\text{ind}} = 1 - \sum_{j} P^2(j \mid t), \tag{2}$$

Table 1: Answers to imposter syndrome questions by medical college students according to the Young Imposter Syndrome scale.

Sl.	Questions	Answer (Yes/No)	Medical college	
			Public (n, %)	Private (n, %)
1	Do you secretly worry that others will find out that you are not as bright and capable as they think you are? (Y1)	Yes	138 (59.2%)	123 (46.1%)
		No	95 (40.8%)	144 (53.9%)
2	Do you sometimes shy away from challenges because of a nagging self-doubt? (Y2)	Yes	137 (58.8%)	105 (39.0%)
		No	96 (41.2%)	163 (61.0%)
3	Do you tend to chalk your accomplishments up to being a fluke no big deal or the fact that people just like you? (Y3)	Yes	114 (48.9%)	127 (47.6%)
		No	119 (51.1%)	140 (52.4%)
4	Do you hate making a mistake, being less than fully prepared or not doing things perfectly? (Y4)	Yes	115 (49.4%)	144 (53.9%)
		No	118 (50.6%)	123 (46.1%)
5	Do you tend to feel crushed even by constructive criticism, seeing it as evidence of your ineptness? (Y5)	Yes	119 (51.1%)	142 (53.2%)
		No	114 (48.9%)	125 (46.8%)
6	When you do succeed, do you think, "phew" I fooled them this time but I may not be that lucky next time? (Y6)	Yes	103 (44.2%)	113 (42.3%)
		No	130 (55.8%)	154 (57.7%)
7	Do you believe that other people (students, colleagues, competitors) are smarter and more capable than you? (Y7)	Yes	113 (48.5%)	158 (59.2%)
		No	120 (51.5%)	109 (40.8%)
8	Do you live in fear of being found out, discovered, or un-masked? (Y8)	Yes	111 (47.6%)	105 (39.3%)
		No	122 (52.4%)	162 (60.7%)

where $p(j|t)$ is the estimated class probability for feature or node t in the decision tree and j is the output data or class. In our study, $j = 2$ is represented as IS = Yes and IS = No. Because MDGI is more robust than the mean decreases of accuracy, it was used to select the important IS parameters [25]. The IS parameter with the highest MDGI value is regarded as the most important feature because it has the greatest impact on predictive performance.

Random forest classifier is employed in this study from the Scikit-learn python package [26]. The model is tuned with the following parameters:

(i) n_jobs: −1. This parameter represents the number of jobs running parallelly. −1 is used to instruct the system to use all available processors in the system, rather specifying the number of jobs.

(ii) criterion: gini. This parameter represents the measure of the quality of the data split. For this study as the measure, Gini impurity is utilized.

(iii) max_features: 0.9. This parameter is used to specify the number of features to be used in the data split mechanism. Here, 90% of the features are utilized in each split.

(iv) max_depth: 4. This parameter points the maximum depth of the tree. In this study, maximum depth of the tree is specified to 4.

(v) eval_metric_name: logloss. This parameter is used to evaluate the model performance. In this model, logloss is used as the evaluation metric.

(vi) validation_type: kfold. This parameter is used to validate the model prediction on new data. In this model, Kfold is used for the validation.

(vii) k_folds: 5. This parameter is used to provide the k value for the kfold validation method. For this model, a K value of 5 is used for the kfold validation measure.

2.5. Neural Network. In our study, we used the artificial neural network (ANN) method to make a primary prediction as part of the ensemble method. The ANN is a mathematical model that replicates actual biological neural aspects to make a computational model that can map input and output [27]. Each neuron works like a basic unit that performs the following equation.

$$y = \max\left(0, \sum_i w_i x_i + b\right), \tag{3}$$

where y is the neuron output, x_i is the neuron output, w_i is the weight, and finally, b is the deviation. Each neuron gives a single output (y) from all input (x_i). All neurons are connected through multilevel architecture [28] where the orientation of input and output is done by employing the following equations:

$$h_i = \max\left(0, W_i.h_{i-1} + b_i\right), \tag{4}$$

$$y = \max\left(0, V.h_L\right), \tag{5}$$

where W_i and V are matrices b_i is the vector that is learned from the parameters of the dataset. L is the number of layers or levels and $1 \leq i \leq L$, $h_0 = x$.

In our case, the ANN input includes normalized values of 9 variables such as institution, age, sex, BMI, smoking status. And the output value is the presence of imposter syndrome (IS = yes or no). All datasets are divided into two parts, training and test set. $K = 5$-fold cross-validation is used as a method of validation. We used logloss value as a measure of performance for ANN [29]. Mathematically logarithmic loss or log loss is defined by

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log,\ p_{ij}, \tag{6}$$

where N is the number of samples, M is the number of possible labels, y_{ij} is the binary indicator for whether or not j is the right classifier for i, and p_{ij} is the model probability assigned to j for instance i.

For this study, mljar's AutoML [30] method is used for employing a neural network that utilizes TensorFlow [31] and Keras [32] for the neural network algorithm. Applied parameters for the neural network are as follows:

(i) n_jobs: −1. This parameter represents the number of jobs running parallelly. −1 is used to instruct the system to use all available processors in the system, rather specifying number of jobs.

(ii) dense_1_size: 32. This parameter represents the dense layer one for this Neural Network model. For this model, 32 neurons are used in the first dense layer of the network.

(iii) dense_2_size: 16. This parameter denotes the size of the second dense layer of the network, where the number of neurons is 16 in this dense layer.

(iv) learning_rate: 0.05. This parameter is used for setting up the learning rate of the model. A learning rate is the measure of change that needs to be updated in response to the estimated error each time the weight of the model is updated.

(v) validation_type: kfold. This parameter is used to validate the model prediction on new data. In this model, Kfold is used for the validation.

(vi) kfolds: 5. This parameter is used to provide the k value for the kfold validation method. For this model, a K value of 5 is used for the kfold validation measure.

(vii) eval_metric_name: logloss. This parameter is used to evaluate the model performance. In this model, logloss is used as the evaluation metric.

2.6. Ensemble Model. In this study, we employed forward stepwise selection techniques for ensemble modelling. We used forward stepwise selection from a library of models to generate a subset of models that, when averaged together, gives relatively better predictive results than other models. In this particular modelling technique, three principles are involved in the process to make a better prediction and reduce the overfitting of the data [33] which are as follows:

2.6.1. Selection with Replacement. Model selection without replacement improves when best models are introduced to the ensemble but quickly decreased the performance of the ensemble as other models are being included in the ensemble. In this case, model selection with replacement provides a better prediction for all performance metrics [34].

2.6.2. Sorted Ensemble Initialization. When ensembles are small, the forward selection can overfit the selection. In this process, instead of starting with the empty ensemble, we sorted the models in the library according to their performance and finally put the best models ANN and RF, in the ensemble.

2.6.3. Bagging of the Ensemble Models. Ensemble learners are prone to overfitting the data. So, we took a sample from the data and trained the models. Once we are done, we took another sample so on. Finally, we bagged ANN and RF model predictions together using a simple average of the predictions.

In this study, for building an ensemble model, a random forest classifier and a neural network are used, and both models are equally weighted. The model is trained with a random forest classifier and then with neural networks and combining their prediction settings; our ensemble model is generated. Figure 2 represents the predictive ensemble model generation in this study.

2.7. Performance Metrics. In this study, we used Kfold cross-validation instead of train test split as the train test split method introduces bias to the model. In order to evaluate our predictive model, we used several measurements such as sensitivity/recall (Sn), accuracy (Ac), precision (Pn), F1-score, logloss, and Matthews correlation coefficient (MCC). The equations for Ac, Sn, Pn, F1-score, and MCC measurements are presented as follows:

$$\begin{aligned}
\text{accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100,\\
\text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100,\\
\text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}},\\
F1 - \text{score} &= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}},\\
\text{MCC} &= \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},
\end{aligned} \tag{7}$$

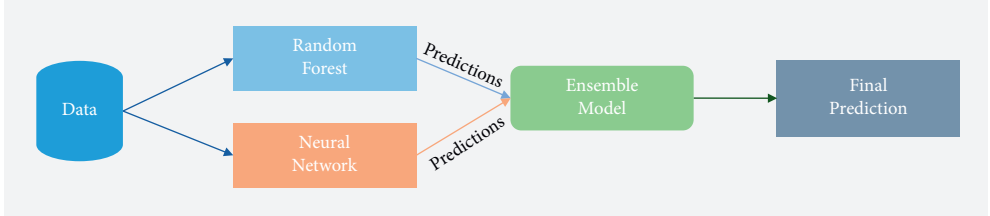

Figure 2: Predictive ensemble model generation using random forest and neural network models.

Table 2: Evaluation metrics for the random forest, neural network, and ensemble models.

	Random forest model	Neural network model	Ensemble model
Accuracy	0.935	0.963	0.964
Precision	1	1	1
Recall	1	1	1
F1	0.872727	0.933333	0.933334
Logloss	0.335263	0.263876	0.206847
AUC	0.971482	0.986442	0.988832
MCC	0.838383	0.905834	0.906743

For all models, accuracy, precision, recall, F1-score, logloss, AUC, and MCC values are measured as evaluation metrics.

where TP, TN, FP, and FN are the number of true positive, true negative, false positive, and false negative, respectively. An MCC coefficient +1 represents perfect prediction when 0 represents random prediction and −1 represents total disagreement between observation and prediction [35]. In addition, we also demonstrated ROC (receiver operating characteristics) and AUC (area under curve).

2.8. Experimental Setup. For this experiment, *Python* programming language is used. The version of python used is 3.6.5 with Anaconda. Keras version 2.4.3, scikit-learn version 0.23.1, tensorflow 2.2.0, and mljar-supervised version 0.8.9 are used for model training and prediction model design. As the operating system, windows 10 version 21H1 (build 19043.1151) 64 bit is used. As hardware, Intel Core i3 6300T with 12 GB RAM and SSD setup is used.

3. Results and Discussion

The experiment shows that the ensemble prediction model provides better results in terms of accuracy, precision, recall, F1, AUC, logloss, and phi coefficient (mcc) metrics than the random forest and neural network's individual prediction model for predicting YIS. In Table 2, the output for all evaluation metrics for all the models is provided.

From Table 2, for all evaluation measures, the neural network performs better than random forest, and the ensemble model performs better than the neural network. For the accuracy of this predictive model, the ensemble model is the highest performer with an accuracy of 96.4%, where individual accuracy for random forest and neural network is 93.5% and 96.3%, respectively. Though there is a very slight improvement in the accuracy measure between the ensemble and neural network model, there is a good amount of improvement in logloss value in the ensemble model over the other two individual models. LogLoss (logloss) value is one of the most important measures for classification problems in terms of probability. The lower the logloss value is, the better the prediction is. This prediction result can be clearer with the confusion matrix represented in Figure 3. From Figure 3, it is seen that, in the confusion matrix of the random forest model, the true positive value is 1, which means it classifies all true positive as true positive values. However, the true negative score is .77, which means this model classifies true negative classes as true negative 77% times and 23% times it classifies false negatives. With the neural network model, this false negative value is decreased to 10% from 23%, and in the ensemble model, these settings are kept and used for the final prediction. In the final prediction model, the false negative predictions are decreased, compromising the false positive prediction of 1%.

From the receiver operating characteristics (ROC) curve of the final ensemble predictive model provided in Figure 4, the predictive model's performance can be observed. This figure contains the ROC curve for positive and negative classes with a micro average of the measures. The area for class 0 and class 1 is 99%, which is a good representation of the model's prediction capability as 100% is the maximum value for this measure.

Feature importance is also calculated for both random forest and neural network models. A heatmap of feature importance is presented in Figure 5. For both models, top valued features are common, and they mostly belong to the questions. Table 3 provides the side-by-side importance value comparison for features pointed out by random forest and neural network models. For the neural network, top feature is Y7 (a student believes that they are less smart than the other people), and for the random forest, Y2 (a student sometimes shy away from challenges because of a nagging

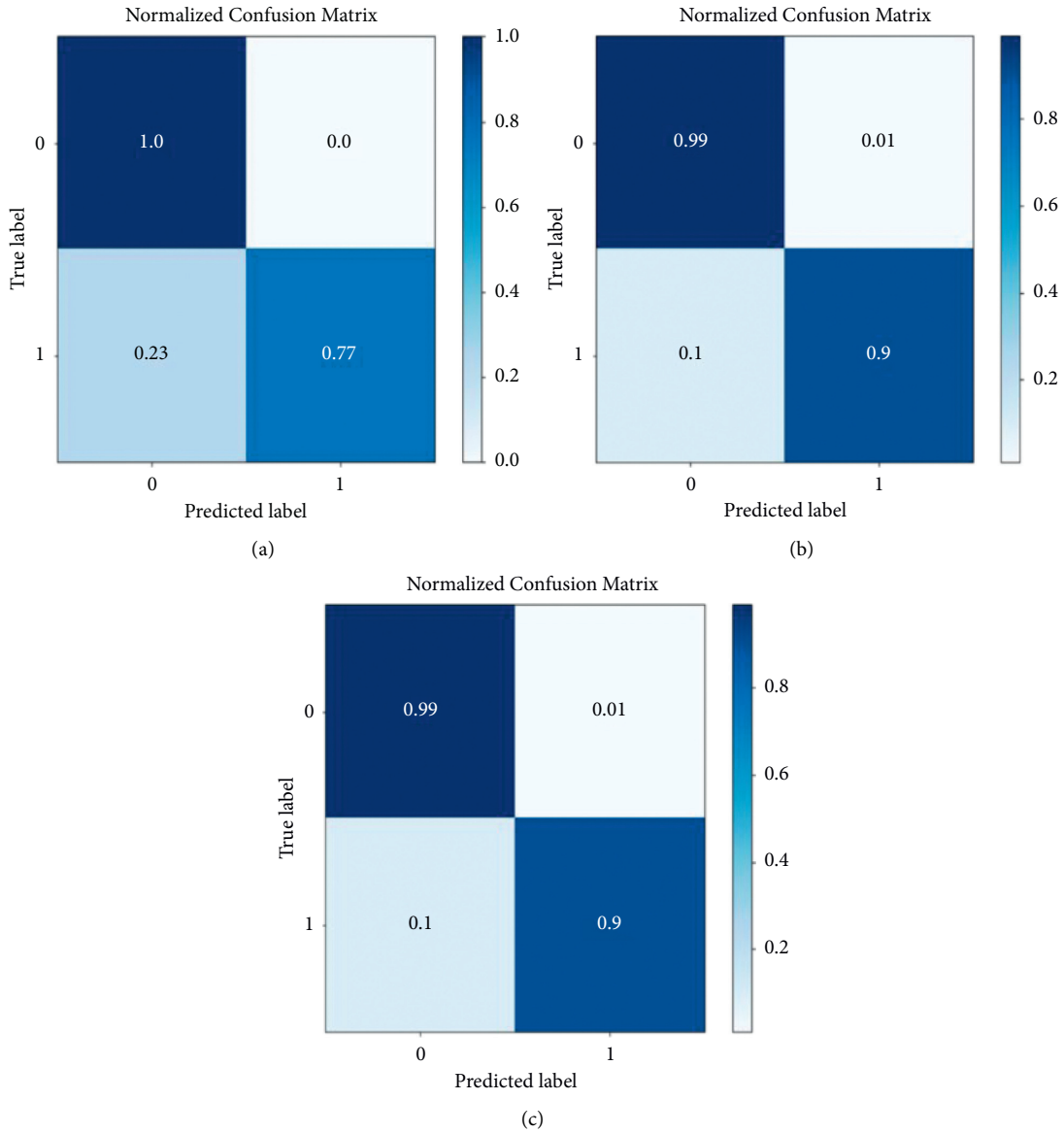

Figure 3: Normalized confusion matrix for random forest, neural network, and ensemble models.

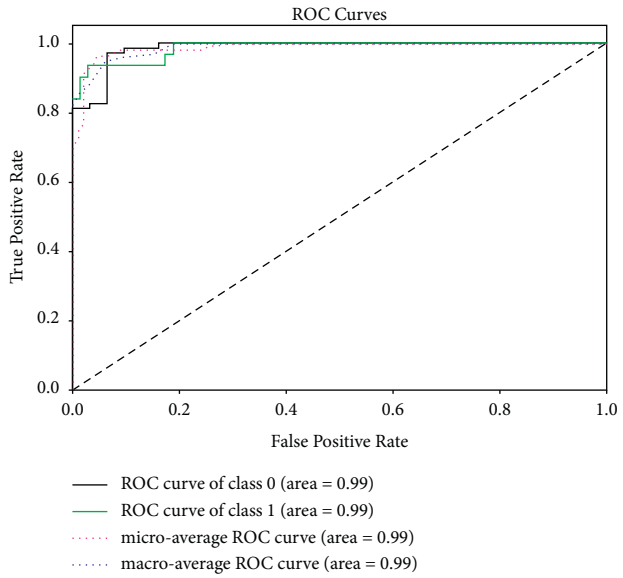

Figure 4: ROC curves for different classes of the dataset generated by the employed ensemble model.

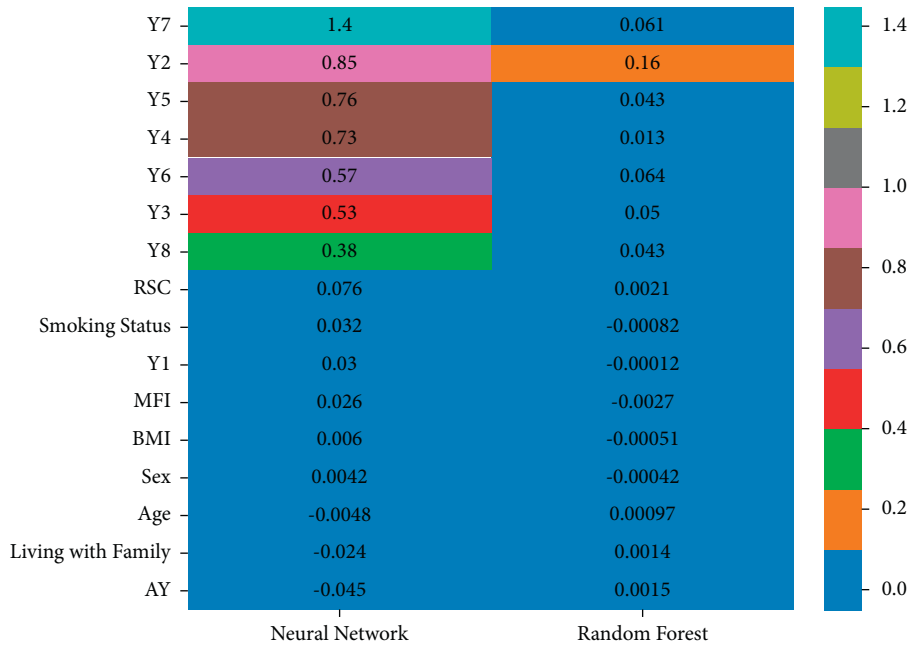

Figure 5: Heatmap of feature importance scores obtained from random forest and neural network model's feature importance method for the features of the young imposter syndrome dataset.

Table 3: Feature importance scores obtained for the features of the young imposter syndrome dataset employing neural network and random forest model.

Feature	Neural network	Random forest
Y7	1.435915476	0.060705199
Y2	0.848385469	0.162305453
Y5	0.758418731	0.042913745
Y4	0.729478374	0.012774768
Y6	0.566908382	0.064326011
Y3	0.531113149	0.050292544
Y8	0.381514902	0.042809194
Reason for study choice (RSC)	0.076035131	0.002130504
Smoking status	0.031637418	−0.000818094
Y1	0.029711063	−0.000117368
Monthly family income (MFI)	0.026247537	−0.002682493
BMI	0.005986669	−0.000505573
Sex	0.00416312	−0.000420911
Age	−0.004779267	0.000966939
Living with family	−0.024055834	0.001383877
Academic year (AY)	−0.045372056	0.001460318

self-doubt) is the top feature, which is the second to top feature for the neural network model. If we compare the top 8 features between neural network and random forest, we can see that the top 8 features of neural network lie between 0.38 and 1.4 whereas, in the case of the random forest, they lie between 0.043 and 0.061. In every case, the neural network assigns a relatively higher importance score to the variables than the random forest algorithm. The main goals of feature selection are to avoid overfitting and improve model selection and also to gain deeper insight into the underlying process [36]. According to the goals of feature selection, better-fitted models give a relatively higher feature importance score than other less-fitted models. In this case, in Table 3, we showed that in every case, the general accuracy of the neural network when predicting the IS is better than the random forest model.

After reviewing the experimental result, it can be deduced that utilizing an ensemble model is better in predicting outcomes rather than utilizing a single machine learning model. This study aimed to generate a predictive model with better predictions for young imposter syndrome, and the ensemble model provides better prediction than two other single models having an accuracy of 96.4% and an AUC score of 98.8%.

4. Conclusion

In this research, we discussed the accuracy of different machine learning algorithms in predicting IS. We also demonstrated that when predicting IS, combined methods such as ensemble learning outperform single machine learning methods such as random forest and neural network. Importance score was used to illustrate how two different machine learning techniques assigned different importance scores to the same variable. The data we used in our work were collected from different medical colleges rather than a single institution to cover most of the medical students studying in Dhaka. The primary aim of this study is to identify a well-fit model in predicting IS. To determine whether a student had IS, we used the YIS scale. We compared the results of each machine learning technique using multiple performance metrics. When compared to individual learners, the ensemble learning model scored higher on all performance matrices.

IS is an emerging mental illness in Bangladesh and requires the immediate attention of researchers. For instance, in order to reduce the impact of IS, identifying key factors responsible for IS is an important step. Machine learning methods can be employed to identify the potential sources responsible for IS. Similarly, determining how each factor contributes to the IS condition among medical students could be a potential future direction.

Data Availability

The data and questionnaire can be found in *Imposter Syndrome Data Set* [37, 38]. Data are available under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0). All the python codes can be found in *Data and Code Repository for Young Imposter Paper for Complexity* [39].

Ethical Approval

Ethical approval was obtained from the Institutional Review Board (IRB) at North South University, Dhaka, Bangladesh. The study objectives were explained to each participant and confidentiality, and anonymity was assured.

Consent

Written consent was obtained from respondents before proceeding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] M. Breeze, "Imposter syndrome as a public feeling," *Feeling Academic in the Neoliberal University*, Palgrave macmillan cham, London, UK, pp. 191–219, 2018.

[2] K. K. L. Mak, S. Kleitman, and M. J. Abbott, "Impostor phenomenon measurement scales: a systematic review," *Frontiers in Psychology*, vol. 10, p. 671, 2019.

[3] E. Hendriksen, "What is impostor syndrome?" *Scientific American*, 2015, https://www.scientificamerican.com/article/what-is-impostor-syndrome.

[4] P. R. Clance and S. A. Imes, "The imposter phenomenon in high achieving women: dynamics and therapeutic intervention," *Psychotherapy Theory Research and Practice*, vol. 15, 1978.

[5] D. M. Bravata, S. A. Watts, A. L. Keefer et al., "Prevalence, predictors, and treatment of impostor syndrome: a systematic review," *Journal of General Internal Medicine*, vol. 35, no. 4, pp. 1252–1275, 2020.

[6] M. Gottlieb, A. Chung, N. Battaglioli, S. S. Sebok-Syer, and A. Kalantari, "Impostor syndrome among physicians and physicians in training: a scoping review," *Medical Education*, vol. 54, no. 2, pp. 116–124, 2020.

[7] J. B. Sullivan and N. L. Ryba, "Prevalence of impostor phenomenon and assessment of well-being in pharmacy residents," *American Journal of Health-System Pharmacy*, vol. 77, no. 9, pp. 690–696, 2020.

[8] A. M. Holliday, G. Gheihman, C. Cooper et al., "High prevalence of imposterism among female harvard medical and dental students," *Journal of General Internal Medicine*, vol. 35, no. 8, pp. 2499–2501, 2020.

[9] J. A. Villwock, L. B. Sobin, L. A. Koester, and T. M. Harris, "Impostor syndrome and burnout among American medical students: a pilot study," *International Journal of Medical Education*, vol. 7, pp. 364–369, 2016.

[10] M. Y. Ikbaal and N. A. S. Musa, "Prevalence of impostor phenomenon among medical students in a Malaysian private medical school," *International Journal of Medical Students*, vol. 6, 2018.

[11] M. A. Qureshi, J. Taj, M. Z. Latif, S. Zia, M. Rafique, and M. A. Chaudhry, "Imposter Syndrome among Pakistani Medical Students," *Annals of King Edward Medical University*, vol. 23, 2017.

[12] K. T. Wang, M. S. Sheveleva, and T. M. Permyakova, "Imposter syndrome among Russian students: the link between perfectionism and psychological distress," *Personality and Individual Differences*, vol. 143, pp. 1–6, 2019.

[13] A. Rosenstein, A. Raghu, and L. Porter, "Identifying the prevalence of the impostor phenomenon among computer science students," in *Proceedings of the Annual Conference on Innovation and Technology in Computer Science Education*, pp. 30–36, ITiCSE, Portland, OR, USA, March 2020.

[14] B. Levant, J. A. Villwock, and A. M. Manzardo, "Impostorism in third-year medical students: an item analysis using the Clance impostor phenomenon scale," *Perspectives on Medical Education*, vol. 9, no. 2, pp. 83–91, 2020.

[15] P. M. Niemi and P. T. Vainiomäki, "Medical students' distress—quality, continuity and gender differences during a six-year medical programme," *Medical Teacher*, vol. 28, no. 2, pp. 136–141, 2006.

[16] A. G. Burstein, S. Loucks, J. Kobos, G. Johnson, R. L. Talbert, and B. Stanton, "A longitudinal study of personality characteristics of medical students," *Academic Medicine*, vol. 55, no. 9, pp. 786-787, 1980.

[17] N. Jahan, S. Islam, and R. Sultana, "Factor scoring and machine learning algorithm to predict student counselling," *International Journal of Engineering and Advanced Technology*, vol. 9, 2019.

[18] G. D. Israel, *Determining Sample Size 1 The Level of Precision*, University of Florida, Gainesville, FL, USA, 1992.

[19] F. Q. Nuttall, "Body mass index: obesity, BMI, and health: a critical review," *Nutrition Today*, vol. 50, no. 3, pp. 117–128, 2015.
[20] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, 2018.
[21] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, 2016.
[22] H. A. Abdi, "Neural network primer," *Journal of Biological Systems*, vol. 2, 1994.
[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, 2001.
[24] G. G. Moisen and T. S. Frescino, "Comparing five modelling techniques for predicting forest characteristics," *Ecological Modelling*, vol. 157, 2002.
[25] M. L. Calle and V. Urrea, "Letter to the editor: stability of random forest importance measures," *Briefings in Bioinformatics*, vol. 12, pp. 86–9, 2011.
[26] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, 1992.
[27] A. Behnood and E. M. Golafshani, "Predicting the compressive strength of silica fume concrete using hybrid artificial neural network with multi-objective grey wolves," *Journal of Cleaner Production*, vol. 202, 2018.
[28] F. Deng, Y. He, S. Zhou, Y. Yu, H. Cheng, and X. Wu, "Compressive strength prediction of recycled concrete based on deep learning," *Construction and Building Materials*, vol. 175, 2018.
[29] B. C. Andrew, "Making Sense of Logarithmic Loss - Datawookie," 2015, https://datawookie.dev/blog/2015/12/making-sense-of-logarithmic-loss/.
[30] Github, "AutoML mljar-supervised," 2021, https://supervised.mljar.com/.
[31] A. F. Gad, *TensorFlow: A Guide to Build Artificial Neural Networks Using Python*, LAP LAMBERT Academic Publishing, Sunnyvale, CA, USA, 2017.
[32] Keras, "Developer guides," 2021, https://keras.io/guides/.
[33] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, 1997.
[34] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble selection from libraries of models," in *Proceedings of the Twenty-First International Conference on Machine Learning*, July 2004.
[35] M. Vihinen, "How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis," *BMC Genomics*, vol. 13, no. Suppl 4, 2012.
[36] Y. Qi, *Random Forest for Bioinformatics*, Ensemble Machine Learning, Princeton, NJ, USA, 2012.
[37] M. Shahjalal, "Distribution of imposter syndrome (IS) among medical students of Bangladesh: a cross-sectional study," *Mendeley Data*, 2021.
[38] M. Shahjalal, M. N. A. Khan, F. M. Mohsin et al., "Distribution of imposter syndrome among medical students of Bangladesh: a cross-sectional study," *F1000Research*, vol. 10, p. 1059, 2021.
[39] M. N. A. Khan, T. B. Sarwar, and M. S. U. Miah, "Data and code repository for young imposter paper for complexity. GitHub," 2021, https://github.com/ping543f/git_version_for_paper_yi.git.