



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Enhancing Graduate Employability Prediction Through Ensemble Learning For Educational Decision Support

Abha Goswami^{1*}, Gagandeep Chawla¹, Charu Gupta²

¹ University Institute of Computing, Chandigarh University, Punjab, 140413, India
Email: goswamiabha7@gmail.com

² Bhagwan Parshuram Institute of Technology, Delhi, 110089, India.

Abstract: In higher education, employability prediction has become an important research domain wherein, universities increasingly employ methods to predict employability and their associated patterns. In this context, ensemble models tend to outshine traditional single models inherently due to their high prediction accuracy, low variance and reduced bias. Also, as per the survey of India Skill Report 2025 in India, Punjab (state) and Chandigarh (UT) have shown lowest grade of employability pattern. Therefore, for the purpose of this study, data is collected through a questionnaire from colleges and universities across Chandigarh and Punjab in India to study the pattern and prediction of employability. The Cronbach's Alpha value for the questionnaire was 0.947 which indicates reliability of the data collected. The proposed methodology uses four machine learning models—Support Vector Machines, Multilayer Perceptron, Random Forest, and K-Nearest Neighbors to predict students' employability. In this study, the ensemble approach provides a strong predictive framework that uses genetic algorithm to optimize learning parameters and improve adaptability. The use of genetic algorithm helps in identifying and discarding suboptimal candidate solutions which further helps in building iterative candidate solution population that shows efficient convergence. The performance of the proposed method is computed using accuracy, precision, recall, F1-score, support, and ROC-AUC score. The proposed model provides 91.8% prediction accuracy on the primary dataset and 85.5% accuracy on an external Kaggle dataset. The findings highlight that the model offers high predictive capability and generalizability for graduate employability prediction.

Keywords: employability, genetic algorithm, ensemble techniques, prediction, Higher education, machine learning.

Introduction

Employability of students is essential for both individual and societal well-being and contributes to economic development. However, graduate employability prediction remains a national challenge in educational analytics, as every year a cohort of graduates emerges from educational institutions. After graduation, students want to have promising careers and expect their institutions to assist them in finding work [1]. As competition is increasing in the market, it is important to detect the features that will affect students employability [2],[3]. Previous studies have extensively employed data-driven approaches in various educational data mining fields [4],[5],[6]. With advancements in technology, researchers have turned more toward predicting employability through machine learning models, including linear and logistic regression, Random Forest (RF) [7], Support Vector Machines (SVM) [8], naïve Bayes, decision trees, K-Nearest Neighbours (KNN) [9], and neural networks such as Multilayer Perceptron (MLP) [10]. These techniques use a prediction model that considers a number of aspects and features for evaluating and predicting students' employability [11],[12]. Different factors that affect employability among students include academic achievement, analytical ability and self-awareness, general knowledge, problem-solving capacity, leadership, entrepreneurial skills, communication ability, self-confidence, mental sharpness, and contextual awareness [13]. Earlier academic performance was used as the main determinant of predicting employability; however, it is now understood to be a multidimensional concept impacted by skills [14]. Also, institutions used to keep track of every student's information from the time of enrolling until they graduated so that they could analyse their data and motivate them to identify areas of weakness in order to acquire required industrial skills [15]. However, such data may be unbalanced and

multivariate [16]. Additionally, much of the existing research on employability prediction is limited in scope—focusing on a single institution, a specific course, or employing only one machine learning method—making the findings less generalizable to different educational contexts [17],[18]. Additionally, very little attention has been given to ensemble approaches that use optimization strategies and validate their performance on independent datasets, while most previous research on employability prediction has focused on increasing the prediction accuracy of standalone models, single institution dataset or traditional ensemble techniques [19],[20]. As a result, current employability prediction models' robustness and generalizability are still restricted, limiting their use across diverse educational environments. Hence, there is a need for a predictive framework that not only delivers consistent performance across many datasets but also achieves high accuracy. To address this research problem, the concept of ensemble models in predicting students' employability has drawn a lot of attention, as these models have the ability to improve accuracy by combining various base learners and ensemble learning techniques, providing more robust and reliable predictions than single-model approaches. Accordingly, the current study creates a strong framework for employability prediction using ensemble learning and Genetic Algorithm (GA) optimization approaches,

This study adopts an approach that blends quantitative and qualitative techniques, wherein qualitative data is transformed into quantitative data. Cronbach's Alpha value of 0.947 indicates excellent internal consistency of the questionnaire. Employment status is taken as the dependent variable, while other features— analytical ability and self-awareness, general knowledge, problem-solving skills, leadership qualities, entrepreneurial capacity, communication ability, self-confidence; mental sharpness, and contextual awareness —are taken as independent variables [13].

In contrast to traditional ensemble techniques, this study proposes a GA based ensemble prediction model, which integrates four distinct models for machine learning: KNN, MLP, SVM, and RF to improve learning adaptability, predictive generalisation and accuracy of the employability prediction model. GA improves the suggested ensemble model's capacity for prediction by effectively determining the optimal learning parameters.

In this paper, base models are selected as they belong to the different learning paradigms, having the capability to represent hidden intricate non-linear associations within the data, balance both interpretability and accuracy, reduce variance and over fitting, and perform reliably in high-dimensional spaces. By using ensemble decision trees to reduce variance, Random Forest is a useful tool for capturing intricate feature relationships. A Support Vector Machine offers maximum class discrimination by performing excellently in high-dimensional datasets. While K-Nearest Neighbours classifies instances by exploring similarity among neighbouring data points, Multilayer Perceptron uses hierarchical feature learning to capture complicated nonlinear patterns. These models collectively suit student employability prediction well because they draw on the strengths of multiple approaches to reach better accuracy and wider generalizability. To show the predictive efficacy and generalizability of the suggested framework, it is assessed through internal validation, comparison with a Logistic Regression baseline, and external validation on a separate Kaggle dataset.

To bridge the identified research gap, this study makes the following contributions: (a) Four base classifiers (RF, SVM, MLP, and KNN), are combined into a GA-optimised ensemble framework to improve prediction of graduate employability. (b) Using multiple performance measures and comparison with baseline machine learning approaches, a thorough assessment is carried out. (c) The primary dataset as well as the external Kaggle dataset are used to assess the adaptability and scalability of the suggested methodology. (d) The proposed framework serves as a practical decision support system for enhancing forecasting of students employability in higher education institutions.

Outline: The motivation behind the suggested study is explained in section 2. Section 3 reviews related research relevant to the proposed model. Section 4 provides a comprehensive presentation of the framework, from compilation of data to comparative evaluation of against conventional models. Lastly, the study summarises its findings and outlines directions for future research in section 5.

Motivation

Many students try to find employment after graduation, but many are unable to secure suitable positions due to the gap between what they study and the practical skills required in the job market. Nowadays, academic credentials alone are not enough to get a job; relevant skills are also equally important [21],[22]. As a result, many students remain unemployed, making this a major concern for both the government and educational institutions [23]. This concern is also consistently emphasised in national reports. According to the India Skills Report and the All India Council for Technical Education Annual Report 2025, about 46% of students are not

employable owing to insufficient necessary industry-relevant skills. Given this situation, there is a need to create intelligent predictive models that can forecast students' employability. These models can help institutes identify those students who might require more training, so that they can prepare students for the workforce. To fulfil this demand, data-driven intelligent systems are becoming more popular in employability prediction, especially in machine learning [24]. Ensemble learning techniques, by combining multiple models, show promising results while addressing different types of students' data. In addition, integration of GA with an ensemble model improves the overall prediction performance by selecting the most relevant features and determining the best network parameters.

Literature Review

Machine learning has gained popularity for predicting students' employability because of its ability in processing vast datasets and uncovering complex trends. This literature review examines the growing need for safe predictive modelling in the field of employment prediction of students. This section reviews previous research papers to inspect the types of machine learning techniques and ensemble methods used, along with their accuracies. Previous research have demonstrated machine learning's efficacy. single and ensemble techniques like artificial neural network, decision trees, SVM, AdaBoost, bagging, Bayes networks, reduced error pruning, extreme gradient boosting, multilayer feed forward, K-means, linear regression, cat boost, light GBM , stochastic gradient descent [25].

Through a close review of the literature, we gain comprehensive overview of the important features that influence students' employability and career accomplishments. The important features include analytical ability and self-awareness, general knowledge, problem-solving skills, leadership qualities, entrepreneurial capacity, academic performance, communication ability, self-confidence, mental alertness and contextual awareness [13]. These factors play a significant role in guiding students on their career paths [26]. In 2024, the studies made use of techniques like SVM, RF, KNN, decision trees, neural networks, linear regression, extreme gradient boosting and AdaBoost for predicting students' employability, with an accuracy range of 70.39%–85.6%. Since the results of a single base model are not impressive, ensemble approaches such as stacking and boosting combined with GA were employed. This integration achieved a peak accuracy of 89.72%, utilising features such as communication skills [27], academic records [28], internship experience [29], confidence [30], alertness, and demographic data [31]. During 2023, the studies employed various machine learning techniques, including RF, SVM, KNN, decision tree, neural network, linear regression, extreme gradient boosting and AdaBoost [32], Bayes networks, CatBoost, for forecasting students' academic performance and employability[33],[34]. These studies concentrate on diverse features like academic scores, personal attributes, family background, communication skills, and employability indicators. When a single base classifier is used, accuracy rates range from 83.33% to 87%, When ensemble methods were applied, prediction accuracy improved significantly to 89.5% [35],[36]. According to the survey of papers published in 2022, the impact of a wide range of features, such as academic achievements, parental income, involvement in extracurricular pursuits, technical and communication-related skills, internship experience, professional certifications, and employability class, on students' employability prediction has been reviewed in the studies [37],[38]. The studies focus on various ML techniques, including SVM, RF, KNN, decision tree, neural network, linear regression, extreme gradient boosting and AdaBoost, Naïve Bayes, artificial neural networks, CatBoost for prediction. These research studies are attaining prediction accuracy rates that range from 75.91% to 88% while using a single machine learning approach. However, the forecast accuracy percentage increased to 89.9% when employing the ensemble approach [42]. The study conducted in 2021 indicates that the research predicts the undergraduate students' career [39] and employment status using ensemble machine learning techniques [40],[41]. The review investigated the features, including academic achievement, technical and analytical ability, communication abilities, critical thinking, problem-solving capabilities, demographic attributes like gender, family income, and parent's education, as well as extracurricular activities such as internships and co-curricular participation. These studies focus on various machine learning techniques for prediction, such as SVM, RF, KNN, MLP, decision tree, neural network, linear regression, extreme gradient boosting and AdaBoost, Naïve Bayes, artificial neural networks, CatBoost, Bagging, with accuracy ranging from 50% to 96.9% [43].[44]. But when the ensemble technique is used, the prediction rises to 98.5%. The studies during the year 2020 analysed academic achievements prediction of students' and placement outcomes using single classifiers with accuracy ranging from 76.74% to 84%, but when combining the optimisation methods GA with ensemble techniques, then accuracy rises to 90% [45]. These studies leverage features such as gender, residential location, cumulative grade point average, field of study, confidence, communication skills, work experience,

parental factors, competition behaviour, and placement status, employing techniques like SVM, RF, KNN, MLP, decision tree, neural network, linear regression, extreme gradient boosting, AdaBoost, Naïve Bayes, artificial neural networks, CatBoosting, Bayes Networks, stochastic gradient descent. An exhaustive examination of existing research from 2020 to 2025 revealed that four models of machine learning, KNN, MLP, SVM, and RF, consistently yield the highest accuracy in prediction. This finding is illustrated through the figure 1.

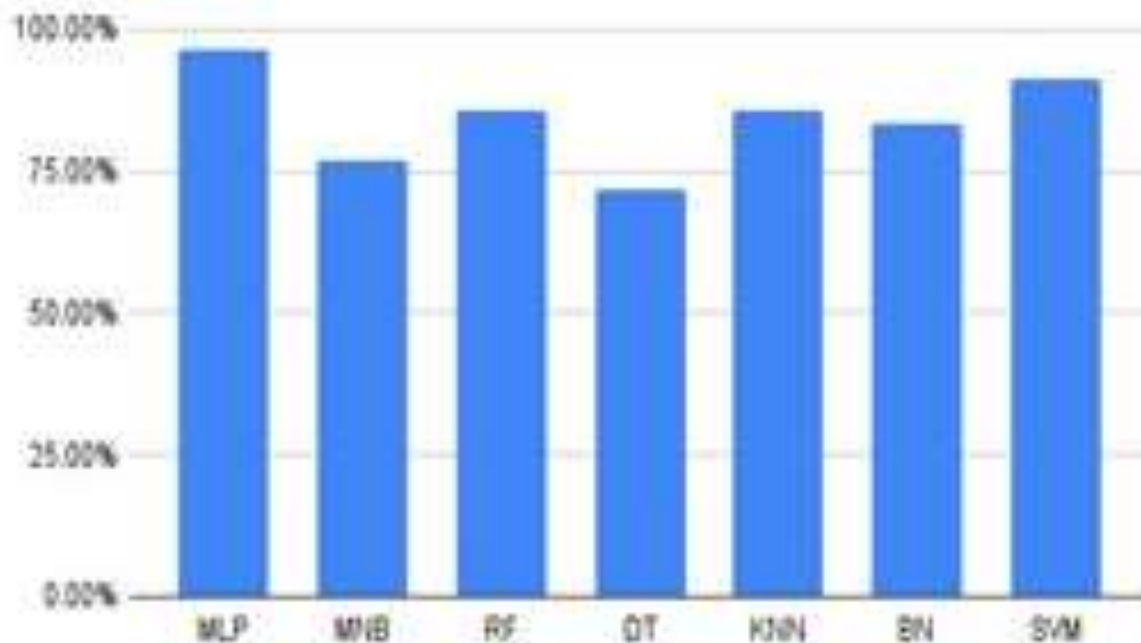


Figure 1: Average accuracy of prediction techniques from 2020 to 2025

Figure 1 depicts the average prediction accuracy that machine learning methods achieved in the previous studies published between 2020 to 2025. It demonstrates that the predictive accuracy of ensemble learning methods is superior to individual base classifiers.

Figure 1 depicts the average prediction accuracy that machine learning methods achieved in the previous studies published between 2020 to 2025. It demonstrates that the predictive accuracy of ensemble learning methods is superior to individual base classifiers.

Methodology

In this section a systematic process is outlined of the proposed work. The pre-processing focuses on cleaning the dataset and applying feature-engineering techniques to prepare it for analysis. Once the data is prepared, four base models-, KNN, MLP, SVM, RF, are trained on 80% of the dataset, with GA incorporated into the MLP model. The accuracy and performance of these models are then evaluated using the remaining 20% of the data. These selected base models are combined using the bagging, boosting, and voting techniques. The resulting proposed ensemble model is assessed using appropriate performance testing metrics, such as ROC, AUC, F1-score, recall, precision and accuracy. Lastly, effectiveness of the predictive ensemble framework is assessed other established models. The structural representation of the research methodology is depicted in figure 2.

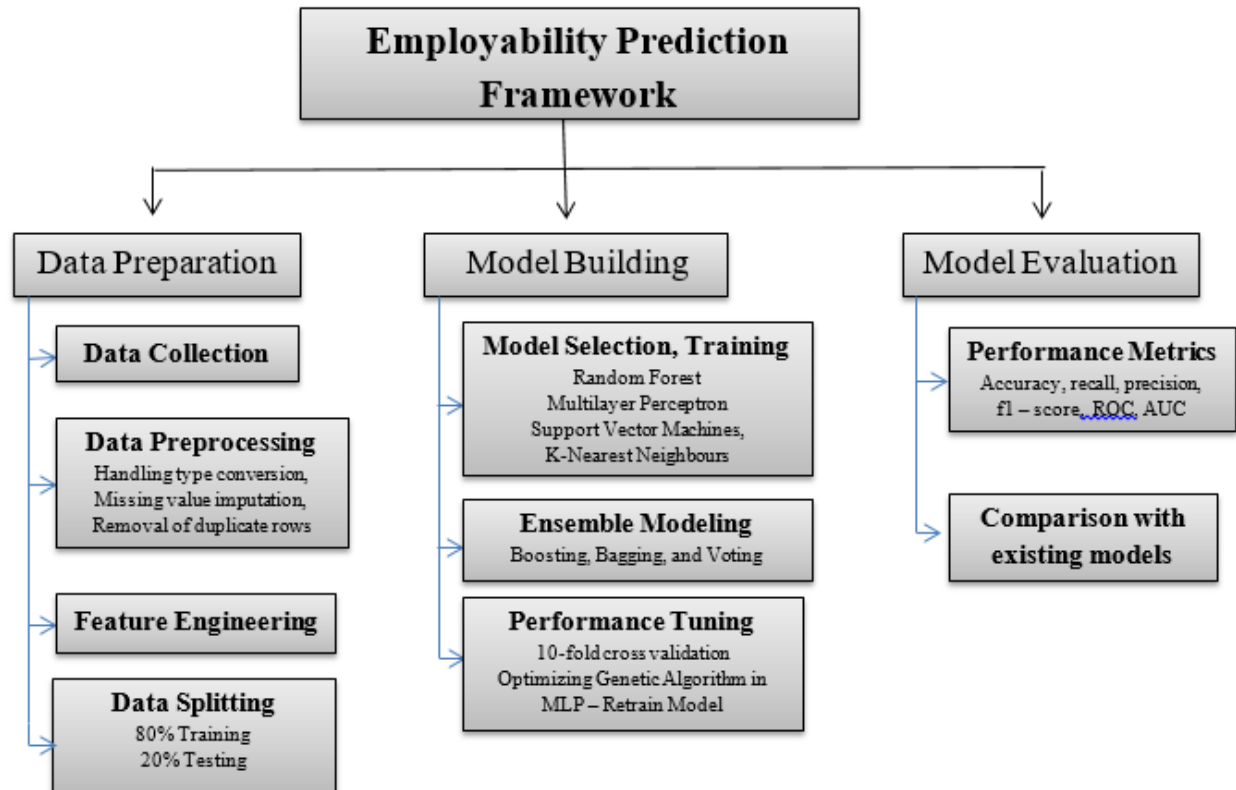


Figure 2: Framework of proposed ensemble model

The entire research framework, encompassing data collection, preprocessing, model creation, ensemble construction, optimization, and model evaluation, is shown in figure 2. The next sections explain each one in detail.

4.1. Data Preparation

Data preparation cleans and convert the raw unstructured data into structured form for model training which includes data collection, data pre-processing.

4.1.1. Data collection

To build an employability prediction model for students, datasets from different modes have been analysed for determining the essential factors influencing the employment status of students. For research, data was gathered through a survey questionnaire approach using Google Forms, The dataset showed the balanced gender representation including replies from a heterogeneous group of undergraduates: 870 males, 929 females, and 5 individuals who identified as other. Each record includes 11 numerical features: written communication, visual communication, active listening, context awareness, technical skills, critical thinking skills, teamwork skills, managerial skills, entrepreneurship skills, self-confidence, and general awareness. All these are independent variables, and one is a dependent variable, i.e., employability status.

4.1.2. Data pre-processing

It acts as a significant in which data undergoes cleaning and is thus prepared to make it suitable for further processing. The surveyed data may contain noise, incomplete values, removal of duplicate rows and may even be in an unsuitable format that prevents machine learning models from using it directly. After that, feature scaling is used to standardize the data.

4.2. Model Building

To train machine learning models on the provided data to identify complex patterns or decision rules for making predictions is referred to as the model building process. Its various components are explained below.

4.2.1. Model selection

After conducting a rigorous literature review, it was found that the four models with the highest frequency in prediction are Multilayer Perceptron, K-Nearest Neighbours, Support Vector Machine, and Random Forest. Therefore, these base classifiers are employed in the proposed ensemble model.

(a) Multilayer Perceptron: The type of this model is a machine learning based neural network architecture, which is used to train complex patterns and non-linear correlations in data [46] It is organised in layers and works well on both organised and unorganised data. This technique used the following equations for prediction. Weighted sum calculations: It is used to convert input data into appropriate form before applying the activation function. This is shown in Equation (1).

$$Z = WX + B \quad (1)$$

Where W is weight matrix, X factor value from previous layer, b is bias term

Activation function (tanh) : It allows the model to discover complex patterns by introducing non linearity into the model. The Equation (2) illustrates this.

$$f(y) = \frac{\exp(y) - \exp(-y)}{\exp(y) + \exp(-y)} \quad (2)$$

Where the total weight of inputs and the bias term given to the neurone are represented by x.

Loss function : It measures how closely actual values match expected values. This is shown in Equation (3)

$$L = -\sum [x \log(\hat{x}) + (1-x)(\log(1-\hat{x}))] \quad (3)$$

Where x denotes true labels, $\hat{x}$ represents predicted outcomes.

Gradient calculation in back propagation: It is employed to adjust the weights in order to improve the predictions. This is shown in the following Equation (4)

$$\frac{dL}{dw_i} = \sum E_i x_{ij} \quad (4)$$

Where L denotes loss function, w_j is weight of jth neuron, $\sum$ ranges from 1 to n (number of training samples), E_i is error for ith sample, x_{ij} is input factor

(b) Random Forest: This ensemble learning framework integrates several decision trees, where each tree is constructed using randomly selected portions of the dataset. The results from all subsets are aggregated for the final prediction. It handles large datasets and high-dimensional factors, with strong resilience to over fitting when adjusted appropriately [47]. This technique used the following equations for predicting the employability. This technique used the following equations.

Decision tree split criteria (entropy or gini impurity): It decides the most effective way, Entropy or Gini Impurity, for splitting data at each decision tree node. Both are used to measure impurity of a node. The following Equations (5) and (6) are used in measurement.

$$\text{Entropy: } H = \sum p_i \log_2 p_i \quad (5)$$

$$\text{Gini Impurity: } G = 1 - \sum p_i^2 \quad (6)$$

Where $\sum$ goes from 1 to c (number of classes), p_i denoted probability of class i

Information gain : It is used to quantify the split's extent, improving classification by reducing entropy. The following Equation (7) is used for splitting.

$$IG(S) = H(P) - \sum \left(\frac{|S_i|}{|S|} \right) H(S_i) \quad (7)$$

Where H (P) is Entropy before split, H (Si) is Entropy of child node, $\sum$ ranges from 1 to n.

(c) Support Vector Machine: It is considered a supervised machine learning method that classifies data by identifying the appropriate hyper plane which best divides data points into different categories [48].

Where ever the data is high-dimensional or the classes are not linearly separable, it performs well. To divide class points of various classes, SVM creates a linear boundary in an n-dimensional vector space and maximises the difference between the linear boundary and the nearest class points, which are known as support vectors.

Feature scaling: It ensures that all factors must be normalised and on the same scale with having 0 as a mean value and 1 as a standard deviation. The following Equation (8) is used

$$X' = \frac{(x - \mu)}{\sigma} \quad (8)$$

Where feature value is denoted by X, mean of factor is shown by μ and σ represents standard deviation.

SVM decision function: Support Vector Classification (SVC) with polynomial kernel is used to find the output class. This is shown in Equation (9) which shows as:

$$F(x) = w^T x + b \quad (9)$$

Where w stands for weight vector, feature vector is denoted by x and bias term by b.

Polynomial Kernel : It transforms data by mapping it into new space using Equation (10).

$$K(y_1, y_2) = (y_1 \cdot y_2 + d)^m \quad (10)$$

Where K (y1, y2) is a kernel function, y1.y2 dot product of factors, d is constant, m is degree which specify complexity of model

(d) K-Nearest Neighbours: It is an instance-based, supervised, non-parametric machine learning technique that classifies or predicts outcomes of an unseen data point by locating its "k" closest neighbours within a training dataset [49]. In essence, it used to predict by considering either the dominant class or the mean value of the nearest neighbours. By comparing the unseen data with the existing data, the model is used to forecast the new information. It uses the following equations for prediction:

Distance Calculation: It measures the spatial distance between two data points. For distance calculation the following Equation (11) is used.

$$d(y, y_i) = \sqrt{\left(\sum (y_i - y_{ij})^2\right)} \quad (11)$$

Where d(y,yi) denotes the spacial distance between a new data point represented by y and yi represents the data used in model training, $\sum$ ranges between 1 to m (number of factors)

Prediction Rule : It is used to predict the output by using mathematical rules and input data. For this purpose, the following Equation (12) is used:

$$y' = \text{mode}(\{y_i \mid x_i \in \text{nearestneighbors}\}) \quad (12)$$

4.2.2 Settings in Training Phase

Two different approaches were employed to build the framework of the proposed ensemble model. The first approach involved creating an ensemble model by combining four different machine learning models: RF, MLP, SVM, and KNN. The configuration shown in Table 1 is used to construct the proposed ensemble model.

Table 1: Configuration of Proposed Ensemble Model

Ensemble	MLP	SVM	KNN	RF
VotingClassifier =Estimators List :Voting = Hard	hidden_layer_sizes=(200, 100): Activation = tanh; solver = adam; max_iter =500; random_state =45,	ProbabilityT rue, kernel='rbf'	n_neighbors=5; n_estimators=10; random_state=42	n_estimators= 10; random_state= 42

In the second approach, a GA is applied to optimise the hyperparameters of the MLP model and then used this optimised MLP within the ensemble. By using these two approaches, we are able to monitor how the proposed model performs under various settings and parameter configurations. The comparison revealed that the GA-based ensemble model achieved higher accuracy than the one without GA. It was chosen because of its capacity to search globally and its efficiency in improving hyperparameters. The suggested ensemble framework's prediction accuracy and reliability are enhanced by using GA to identify the optimal learning parameters, as the MLP is sensitive to the selection of hyperparameters. The parameters used in this approach is outlined in Table 2.

Table 2: Configuration of proposed GA-based Ensemble Model

GA	MLP Classifier	Gradient Boosting Classifier	AdaBoost Classifier	Random Forest Classifier
Generations=30; Crossover rate=.7; Mutation rate=.1; Population size=20	Max Iterations=1000; Learning rate=.01; Hidden layer sizes=(10,)	Base estimator=MLP; Estimators=50; Learning rate=.01; Random state=12	Base estimator=SVM; Estimator=50; Random state=38	Estimators=100; Max depth=10; Random state=22

4.2.3 Reliability checking

To ensure the internal consistency of the questionnaire-based dataset, Cronbach's Alpha is calculated. The computed Alpha value is 0.947, which indicates exceptional internal consistency.

To comprehend the distribution and central tendency of the employability-related attributes, descriptive statistical analysis is conducted. As we have eleven attributes in this study, metrics like mean, median, mode, standard deviation (Std), minimum (Min), first quartile (Q1), third quartile (Q3), and maximum values are calculated using Python-based statistical libraries such as Pandas, NumPy, and SciPy. All these values for each attributes are shown in table 3.

Table 3: Statistical Summary of Employment Skill attributes

Skill	Mean	Median	Mode	Std Dev	Min	Q1	Q3	Max
A1) Written communication	3.51	4	4	1.29	1	3	5	5
A2) Visual communication	3.52	4	4	1.24	1	3	4	5
A3) Active listening	3.73	4	4	1.30	1	3	5	5
B. Context Awareness Skills	3.59	4	4	1.20	1	3	5	5
C1) Technical skills	3.40	4	4	1.24	1	3	4	5
C2) Critical thinking skills	3.59	4	4	1.20	1	3	5	5
C3) Teamwork skills	3.85	4	5	1.31	1	3	5	5
C4) Managerial skills	3.70	4	4	1.27	1	3	5	5
C5) Entrepreneurship skills	3.45	4	4	1.22	1	3	4	5
C6) Self-confidence	3.73	4	4	1.32	1	3	5	5
C7) General Awareness	3.65	4	4	1.27	1	3	5	5

The findings overall indicate areas that must be targeted for skill enhancement. It shows that the students' rating in teamwork, active listening, and self-confidence is higher than entrepreneurship and technical skills. So students must concentrate more on technical competencies rather than solely on theory-based lectures.

4.2.4. Ensemble Modeling

It is a meta-level machine learning strategy that focuses on improving prediction performance by merging the results from other models. In machine learning, ensemble methods frequently yield better results than a single model. The most popular ensemble methods are the following:

Bagging

It is an ensemble learning technique. It aggregates outputs generated from various learners trained on various subsets of the original dataset to improve the precision and effectiveness of algorithms.

Boosting

To minimise the training errors, boosting method creates a strong predictive model by aggregating several low-performing models. Boosting assigns more weight to the instances that are misclassified by the weak learners. Over the repeated iterations, the model increasingly focuses on misclassified instances, resulted in the improved accuracy and efficiency.

Voting

One type of ensemble method that integrates the outcomes of various models is the voting ensemble method. By integrating the skills and knowledge of various experts, the voting ensemble approach can be used to produce forecasts that are more accurate than any one model.

4.2.4 Performance Tuning

Performance tuning involves modifying model parameters to increase accuracy, prevent over fitting, and improve overall model performance. To accomplish this, two techniques have been employed: cross-validation and the use of a GA in the MLP machine learning model. The details are listed below.

Cross-Validation

Using a 10-fold stratified cross-validation approach, a reliable performance evaluation is provided to fit each model on different partitions of the dataset. In this technique, the dataset splits into 10 equal parts, of which nine parts are used for the model's training and the tenth part for evaluation. This procedure is carried out across 10 cycles to ensure reliability.

GA in MLP

It is used to optimise hyper parameters in order to improve model performance through effective search space exploration. GA is used to iteratively develop the optimum parameter combinations through mutation and crossover. This hybrid optimization method improves accuracy of generalization and prediction by maintaining a balance between exploration and exploitation. To enhance the performance of MLP, GA is a suitably effective approach for adjusting the weights and biases of the MLP.

4.2.5 Proposed Work

The pseudocode of the proposed ensemble architecture is shown in Algorithm 1 to demonstrate its functioning stepwise. The proposed ensemble hybrid model's sequential implementation is described in figure 3.

```

Input:
D(1) – primary dataset
D(2) – external dataset
 $X \in \mathbb{R}^{(n \times d)}$  – feature matrix (d = 11)
Y – target variable
n, m – number of samples
(XTR, YTR), (XTS, YTS) – train/test split
TP, FP, TN, FN – True/ False positives, True/ False negatives
N – population size, G – generations
μ – mutation rate, c – crossover rate
LRi – learning rate, fi – fitness
Models - MLR, MMLP, MSVM, MKNN, MRF
Final Ensemble Model - MF
Output:
YP – predicted output

Begin

D(1) =  $\{(x_i, y_i)\}_{i=1}^n$  // Initialisation
D(2) =  $\{(x_j^{ext}, y_j^{ext})\}_{j=1}^m$ 

for k=1 to 2 do
D(k) = D(k) \ {Name} // Data Preprocessing
X(k) = NumericConvert(D(k))
X(k) = MedianImpute(X(k))
D(k) = RemoveDuplicates(D(k))
X(k) = D(k) \ Y(k), Y(k) = employability
End for
(XTR, XTS, YTR, YTS) ← Split(X{(1)}, Y{(1)}, 0.8) // Dataset Partition

LRi ~ U(0.0001, 0.1), i = 1, ..., N // Initialisation of GA

For g = 1 to G : // Optimisation of GA

```

```

 $f_i = \text{Accuracy}(\text{MLP}(LR_i), X_{\{TR\}}, Y_{\{TR\}})$ 
 $(P_1, P_2) = \arg \max f_i$ 
 $child = \frac{P_1 + P_2}{2}$ 
 $child = child + U(-0.01, 0.01)$ 
 $LR_i \leftarrow clip(child, 0.0001, 0.1)$ 
End for

 $LR_{BS} = \arg \max_i f_i$  // Optimal Learning Rate

 $M_{LR} = \text{LogisticRegression}$  (Baseline Model) //Model Construction
 $M_{MLP}, M_{SVM}, M_{KNN}, M_{RF}$  (Base Models)

 $M_k \leftarrow \text{Train}(X_{TR}, Y_{TR}), \forall k \in \{RF, MLP, SVM, KNN\}$  //Model training

 $M_{Bag} = \text{Bagging}(M_{KNN})$  //Ensemble construction
 $M_{GB} = \text{GradientBoost}(M_{\{MLP\}}(LR_{\{BS\}}))$ 
 $M_{Ada} = \text{AdaBoost}(M_{SVM})$ 
 $M_F = \text{Voting}(M_{Bag}, M_{GB}, M_{Ada}, M_{RF})$ 

 $Y_{pred}^k = M_k(X_{TS})$  //Prediction on Test set

 $Acc_k = \frac{TP_k + TN_k}{TP_k + TN_k + FP_k + FN_k}$  // Performance Metrics
 $Prec_k = \frac{TP_k}{TP_k + FP_k}$ 
 $Rec_k = \frac{TP_k}{TP_k + FN_k}$ 

 $\Delta_k = Acc_F - Acc_{LR}$ 
If  $\Delta_k > 0 \Rightarrow M_{\{F\}}$  Outperforms baseline
 $\Delta_k = Acc_{GA} - Acc_F$ 
If  $\Delta_k > 0 \Rightarrow GA$  improves performance
 $Y_{\{pred\}}^{\{ext\}} = M_{\{F\}}(X^{\{ext\}})$  // External Validation
 $Accuracy_{\{ext\}} = \frac{TP_{\{ext\}} + TN_{\{ext\}}}{Total_{\{ext\}}}$ 
 $Acc_{\{ext\}} = \frac{TP_{\{ext\}} + TN_{\{ext\}}}{total_{\{ext\}}}$ 
 $\Delta_{\{lit(i)\}} = Acc_{\{F\}} - Acc_{\{lit(i)\}}$ 
 $\Delta_{\{lit(i)\}} > 0 \rightarrow$  Proposed framework outperforms literature

Return( $Acc_{\{F\}}, Acc_{\{LR\}}, Acc_{\{ext\}}, Acc_{\{lit\}}, F1, Precision, Recall$ )
End

```

Figure 3: Workflow of the Suggested GA-based ensemble model

The proposed GA-based ensemble model's architecture is depicted in figure 3. It incorporates data preprocessing, GA-based optimization, ensemble model construction, and evaluation through internal and external validation.

4.3. Model Evaluation

After training the model, its evaluation is a fundamental phase in determining suggested framework capability for forecasting students' employability. The employability status (employed/unemployed) of every student in the testing set is predicted by the model. The performance of an ensemble model is measured using different

performance indicators like accuracy, F1-score, recall and precision. Additionally, receiver operating characteristics (ROC) curves, and area under the curve (AUC) are used to assess the quality and reliability of the predictions made by the ensemble model.

4.3.1. Performance Metrics

Models are used to forecast the dependent or target variable in the testing dataset. Test data and classifiers are used to assess their performance. Model efficacy is measured by contrasting predicted and actual dependent variable outcomes. The following metrics are applied to assess the effectiveness of a model.

Confusion Matrix: It is the evaluation tool used to compare predicted and actual outcomes. For a classification problem, it shows the count outcomes like true positives, true negatives, false positives, and false negatives, abbreviated as TP, TN, FP, FN, respectively. The total instances for which a positive prediction is accurately identified are known as TP, while the total instances for which a negative prediction is misclassified are known as FP. Negative cases that are accurately classified as negative are referred to be TN. Positive cases that are mistakenly classified as negative are referred to as FN. The confusion matrix is depicted in figure 4.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 4: Confusion Matrix

The confusion matrix's structure, which is used to assess classification performance, is shown in figure 4. It uses true positives, true negatives, false positives, and false negatives for distribution of correctly and incorrectly classified instances.

Accuracy: This metric is a standard measure used to assess the accuracy of the classifier. It determines the portion of data points that are accurately predicted out of all data points. This is shown in Equation (13)

$$A = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (13)$$

Precision: It determines the fraction of positive predictions that are actually correct. Equation (14) illustrates this as follows:

Recall/ TPR (True Positive Rate): It represents the capability of the model to recognize true positive cases. This is illustrated in Equation (15)

F1 – Score: This metric measures the average of precision and recall using the harmonic mean. For this Equation (16) is used:

FPR (False Positive Rate): It measures the extent to which negative samples are incorrectly identified as positive. The Equation (17) of FPR is following:

ROC Curve: It plots TPR and FPR at various threshold levels to show how well a model performs. It uses the following Equation (18).

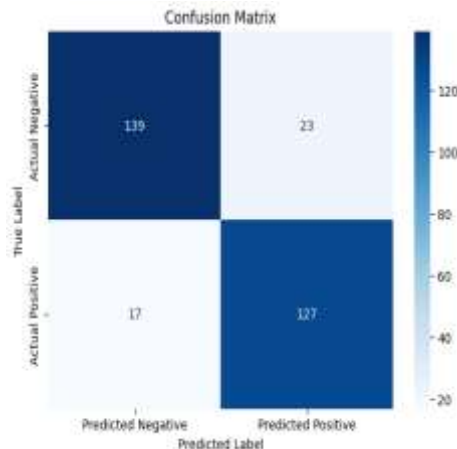
AUC: It indicates the model's ability to distinguish between classes. The Equation (19) of AUC is given below:

Results

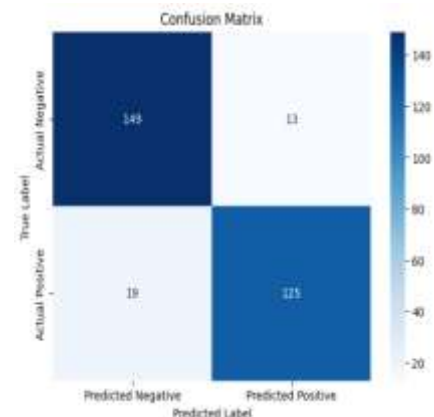
This study's result section is organised to present a detailed examination of outcomes. Evaluation metrics such as the confusion matrix, ROC-AUC, F1-score, precision, recall, support, and accuracy are used to evaluate the predictive performance of the proposed ensemble model. The values of evaluation metrics for each model are shown below.

4.3.2.1 Analysis of Confusion matrix and ROC curves

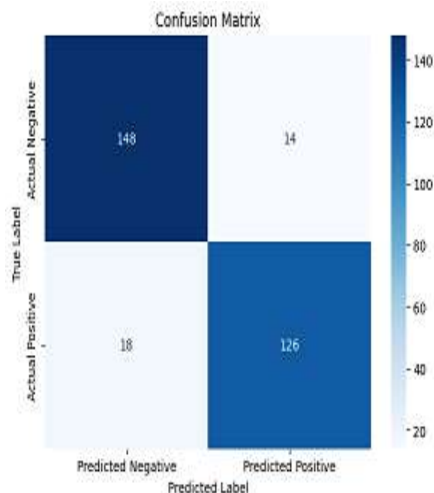
Confusion matrix of machine learning models used in the proposed ensemble model is following:



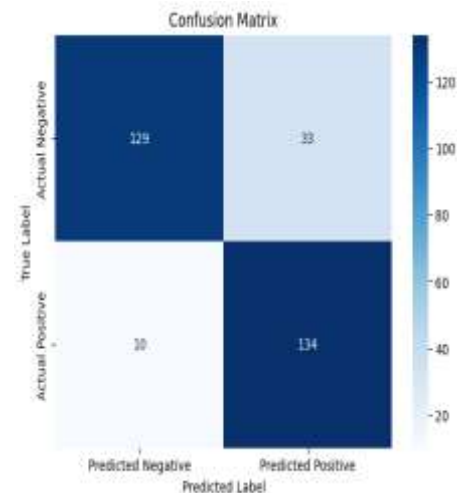
(a). MLP – Confusion Matrix



(b). RF – Confusion Matrix



(c). SVM – Confusion Matrix



(d). KNN – Confusion Matrix

Figure 5: Confusion matrix of four machine learning models

The comparison of four base classifiers is shown in figure 5. It shows the performance of each base classifier before it is included in the suggested ensemble model.

In figure 5, for all four models, the TP are between 125 and 134, while the TN are between 129 and 149. Also, FP range from 17 to 19 and FN range from 13 to 33. These results indicate that the predictions often remain within a steady and constant range across all models.

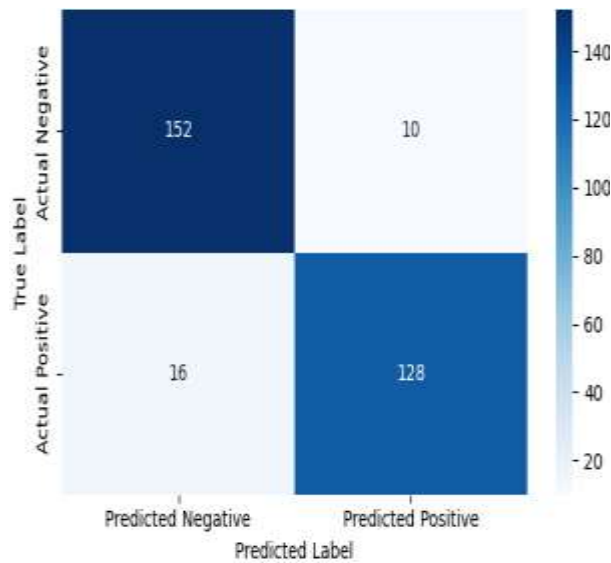


Figure 6: Confusion Matrix of Proposed Ensemble framework

The confusion matrix for the suggested ensemble model is shown in figure 6. It states that the suggested model's performance in classifying employable students from non-employable students has improved. Confusion matrix of a suggested ensemble model is depicted in figure 6. By correctly classifying both classes, employable and non-employable, the model achieves robust predictive performance.

Table 4: Numerical Interpretation of Confusion matrices

Model	TN	FP	FN	TP	Accuracy(%)
MLP	139	23	17	127	86.9
RF	149	13	19	125	89
SVM	148	14	18	126	89
KNN	129	33	10	134	86
Ensemble	152	10	16	130	91.8

The table 4 shows that the ensemble model outperforms all four base classifiers. The ensemble model yields more accuracy with reduced classification errors. Furthermore, ROC – AUC values are used to assess the models' predictive performance.

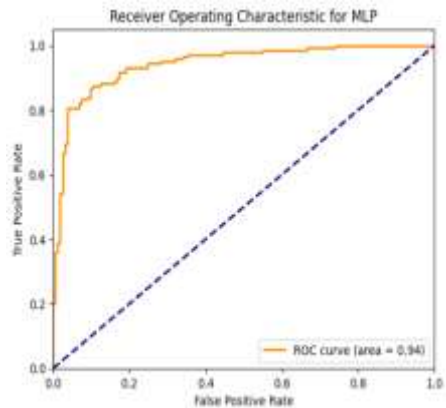


Figure 7(a). MLP – ROC Curve

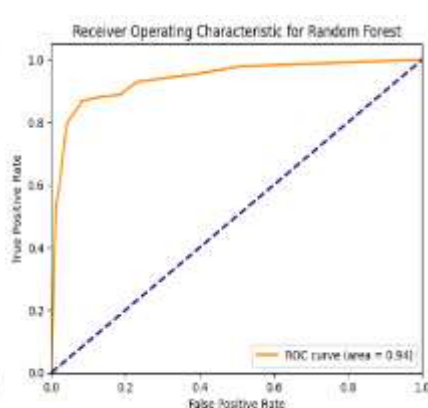


Figure 7(b). RF – ROC Curve

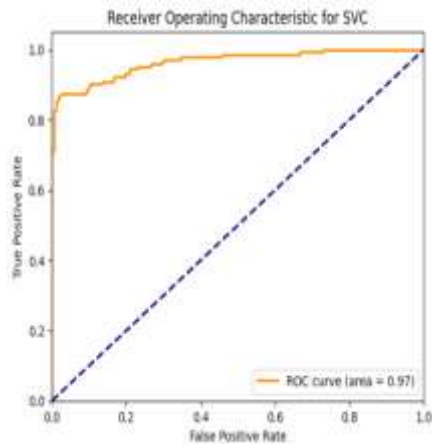


Figure 7(c). SVM – ROC Curve

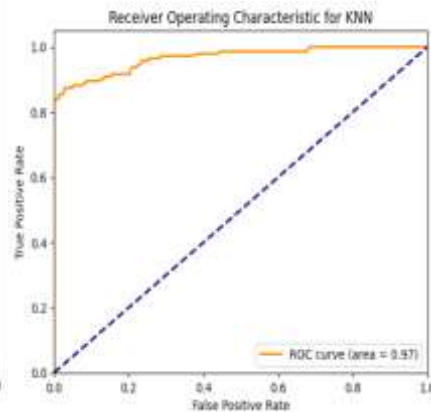


Figure 7(d). KNN – ROC Curve

Figure 7: ROC-AUC of four machine learning models

The ROC curves and AUC values of the four base classifiers are contrasted in figure 7. It demonstrates their capacity to distinguish between students who are employable and those who are not.

Figure 7 shows that all four models have ROC-AUC values between 0.94 and 0.97. This range reflects strong prediction power, where TP rates consistently remain much higher than FP rates.

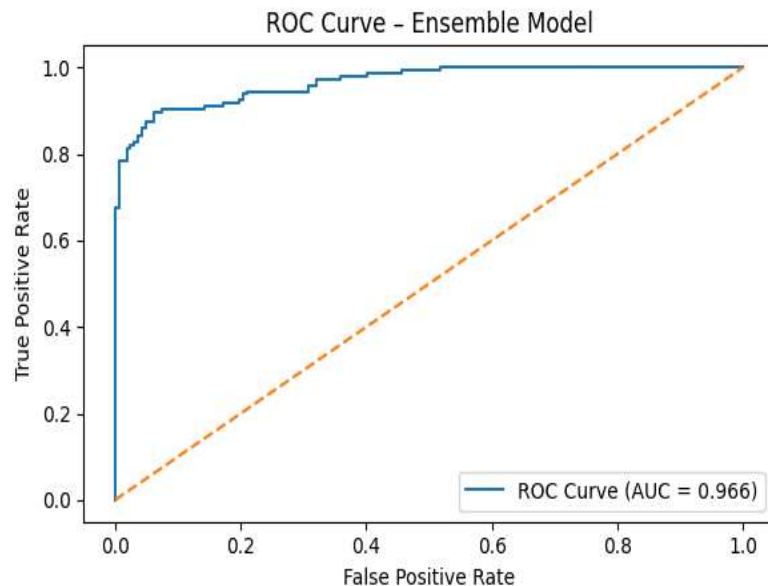


Figure 8: ROC Curve of a Proposed Ensemble Framework

Figure 7 shows that all four models have ROC-AUC values between 0.94 and 0.97. This range reflects strong prediction power, where TP rates consistently remain much higher than FP rates.

Figure 8. ROC Curve of a Proposed Ensemble Framework

The suggested ensemble model's ROC curve is shown in figure 8. Excellent prediction capacity is demonstrated by a high AUC value.

The ensemble model shows robust predictive performance, which is better than random categorization, as seen by the ROC curve in figure 8. This indicates how well the model can differentiate between students who are employable and those who are not.

4.3.2.2 Comparison with existing models

This section compares the proposed ensemble, , framework's performance from three distinct evaluation approaches: internal validation, comparison with existing literature, and external validation.

(a) Internal Validation

Internal validation of proposed ensemble model is done using two approaches:

(i) Benchmarking against a Logistic Regression (LR) Baseline

Logistic Regression is established as the statistical baseline model. All other models used in the proposed ensemble framework are developed using Python, and their results are contrasted with the said baseline model. The code generates the performance comparison table, which shows each model's performance in relation to the baseline in table 5.

Table 5: Performance Comparison of proposed ensemble model with Logistic Regression Baseline Model

Model	F1-Score(%)	Precision(%)	Recall(%)	Accuracy(%)
RF	90	89.1	89.5	89.5
MLP	90.1	90.4	90.2	87.9
SVM	89.3	89.2	89.3	89.5
KNN	90.1	90.2	90.2	86.9
LR Baseline	88.8	88.9	88.9	85.9
GA optimised Model	91.5	91.8	91.5	91.8

Table 5 shows that conventional models such as RF, SVM, MLP, and KNN perform better but still exhibit class-wise discrepancies. While the GA-Optimized model attains the highest and most steady scores for both classes (employable and non-employable), resulting in the highest overall accuracy of 91.8%. This shows that optimizing hyperparameters by using GA improves ensemble model performance compared to the baseline and other traditional models.

(ii) Improvement in model performance via GA based Optimisation

GA is used to refine the MLP classifier's hyperparameters inside the ensemble framework in order to further improve model performance. Table 6 displays the findings, which demonstrate that the GA-optimized model attains higher accuracy along with improved generalization.

Table 6: Comparative Analysis of Base and GA based Ensemble Model

Model	Classes	F1-Score	Precision	Recall	Accuracy
RF	0	.90	.89	.92	89.5%
	1	.89	.91	.87	
MLP	0	.87	.89	.82	86.9%
	1	.86	.85	.86	
SVM	0	.90	.89	.88	89.5%
	1	.89	.90	.88	
KNN	0	.86	.93	.80	85.9%
	1	.86	.80	.93	
Proposed Model	0	.92	.90	.94	91.8%
	1	.91	.93	.91	

From the comparison table 6, it depicts that proposed ensemble approach, attains 91.8% accuracy when optimised using a GA, which makes it the top-performing model.

4.3.2. Performance comparison with existing approaches

The proposed ensemble prediction model is compared with existing models using performance metrics as illustrated in Table 7.

Table 7: Performance Comparison with existing models

Model	F1-Score	Precision	Recall	Accuracy
[21](2024)	.89	.90	.89	89%
[36](2020)	.81	.81	.80	85%

From the comparison presented table 7, it can be observed that the suggested framework exhibits comparable predictive performance with existing GA-based methods. As the studies used for comparison cannot be directly comparable with the suggested framework since they are using different datasets, features, experimental setups, optimization techniques, etc. Specifically, the current study uses Genetic Algorithm (GA) to optimise the MLP hyperparameters inside an ensemble learning framework, whereas other studies only used GA for feature selection or feature analysis. Therefore, rather than proving the superiority of the suggested model, the comparison is meant to offer a qualitative standard.

4.3.3. External validation of proposed ensemble framework

Additionally, an external dataset from Kaggle [50] was used for validation in order to assess dataset independence and investigate the stability of the suggested model. To evaluate the learned ensemble framework's capacity for generalization, it was deployed straight to this unseen dataset without any retraining. The experimental findings show that the suggested ensemble model performed 85.5% accurately on the external dataset. The suggested methodology demonstrated competitive prediction performance even if the external dataset is different from the primary dataset in terms of feature composition and data characteristics. This performance confirms the model's reliability and practical value across different real-world recruitment settings, showing that it keeps competitive and consistent prediction ability beyond the initial training data.

Conclusion and future work

This study examines how educational analytics approaches, GA and ensemble models in predicting the computer science students' employability. It proposes an ensemble model which incorporates a wide range of academic and non-academic factors to capture hidden patterns and trends, reducing variance and improving generalization by outperforming individual classifiers in all evaluation metrics. A GA is integrated into the ensemble model, where it optimizes learning values through processes like crossover, mutation and selection. These processes determine the optimal set of parameters, producing the most accurate performance and improving the model's predictive accuracy. The outcomes of the study can help educational institutions redesign their training programs and course content to better prepare students for careers in computer applications. Higher educational institutions could plan their strategies for students' careers by identifying students who may need extra support in skill enhancement and taking appropriate actions for the benefit of students. They can also plan their curriculum and courses that can meet market demands. The findings support the proposal of a dynamic predictive ensemble employment model that can accurately predict students' employability. For future research, this ensemble model for predicting students' employability can be generalized to other academic disciplines beyond computer science. Secondly, responses can be gathered from a wider geographic area by including states other than Punjab and Chandigarh. Additionally, deep learning and advanced techniques for feature engineering can be explored to further improve the ensemble model's predictive capability.

Declarations

- Funding: No Funding
- Availability of Data: The researchers gathered the first-hand data in research which may be available to others upon reasonable request.

Reference

[1] Ayala Calvo, Juan Carlos, and Guadalupe Manzano García. "The influence of psychological capital on graduates' perception of employability: the mediating role of employability skills." *Higher Education Research & Development* 40.2 (2021): 293-308.

- [2] Fearon, Colm, et al. "Graduate employability, skills development and the UK's Universities Business Challenge competition: a self-determined learning perspective." *Studies in Higher Education* 45.6 (2020): 1280-1297.
- [3] Nair, Prashant R. "Increasing employability of Indian engineering graduates through experiential learning programs and competitive programming: Case study." *Procedia Computer Science* 172 (2020): 831-837.
- [4] Romero, Cristobal, and Sebastian Ventura. "Educational data mining and learning analytics: An updated survey." *Wiley interdisciplinary reviews: Data mining and knowledge discovery* 10.3 (2020): e1355.
- [5] Ahmad, Kashif, et al. "Data-driven artificial intelligence in education: A comprehensive review." *IEEE transactions on learning technologies* 17 (2023): 12-31.
- [6] Shafiq, Dalia Abdulkareem, et al. "Student retention using educational data mining and predictive analytics: a systematic literature review." *Ieee Access* 10 (2022): 72480-72503.
- [7] Kumar, A. Durga Praveen, et al. "Predicting Student Employability Using Machine Learning: A Comparative Study of Classification Algorithms." *International Conference on Computer Science and Communication Engineering (ICCSCE 2025)*. Atlantis Press, 2025.
- [8] Jayachandran, Suja, and Bharti Joshi. "Customized support vector machine for predicting the employability of students pursuing engineering." *International Journal of Information Technology* 16.5 (2024): 3193-3204.
- [9] Qi, Feng, and Maixia Liu. "An AI-driven reinforcement learning framework for graduate employability prediction and career decision support." *Discover Artificial Intelligence* (2026).
- [10] Adawadkar, Kalyani, and Vaibhav Gandhi. "Enhancing employability prediction using stack ensemble learning." *Parul University International Conference on Engineering and Technology 2025 (PiCET 2025)*. Vol. 2025. IET, 2025.
- [11] Khodeir, L. M., & Nessim, A. A. (2020). Changing skills for architecture students' employability: Analysis of job market versus architecture education in Egypt. *Ain Shams Engineering Journal*, 11(3), 811–821. <https://doi.org/10.1016/j.asej.2019.11.006>
- [12] Cortellazzo, L., Bonesso, S., Gerli, F., & Batista-Foguet, J. M. (2020). Protean career orientation: Behavioral antecedents and employability outcomes. *Journal of Vocational Behavior*, 116, 103343. <https://doi.org/10.1016/j.jvb.2019.103343>
- [13] Goswami, A., Chawla, G., & Gupta, C. (2024). Ensemble model analysis for assessing features influencing students' employability in higher educational institutes. In *Proceedings of the International Conference on Computing and Communication Networks* (pp. 603–618). Springer.
- [14] Moumen, A., Bouchama, E. H., & El Idirissi, Y. E. B. (2020). Data mining techniques for employability: Systematic literature review. In *Proceedings of the IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)* (pp. 1–5).
- [15] Dubey, A., & Mani, M. (2019). Using machine learning to predict high school student employability: A case study. In *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 604–605).
- [16] Thakar, P., & Mehta, A. (2017). A unified model of clustering and classification to improve students' employability prediction. *International Journal of Intelligent Systems and Applications*, 9(9), 10–17.
- [17] Mishra, T., Kumar, D., & Gupta, S. (2016). Students' employability prediction model through data mining. *International Journal of Applied Engineering Research*, 11(4), 2275–2282.
- [18] Jantawan, B., & Tsai, C.-F. (2013). The application of data mining to build classification model for predicting graduate employment. *International Journal of Computer Science and Information Security*, 11(9), 1–6.
- [19] Almutairi, M. M., & Hasanat, M. H. A. (2018). Predicting the suitability of IS students' skills for the recruitment in Saudi Arabian industry. In *Proceedings of the National Computer Conference (NCC)*. <https://doi.org/10.1109/NCG.2018.8593016>
- [20] Baffa, M. H., Miyim, M. A., & Dauda, A. S. D. (2023). Machine learning for predicting students' employability. *UMYU Science Journal*, 2(1), 1–8. https://doi.org/10.56919/usci.2123_001
- [21] Thakar, P., & Mehta, A. (2024). Cluster model for parsimonious selection of variables and enhancing students' employability prediction. *arXiv Preprint*. <https://arxiv.org/abs/2407.16884>
- [22] Malika, S. M. M. (2024). Data mining for students' employability prediction. *International Journal of Novel Research and Development*, 9(7), 45–56.
- [23] Deshmukh, A. S., & Raut, A. B. (2023). Students' employability concern for colleges and universities. In *ICT with Intelligent Applications: Proceedings of ICTIS 2022* (Vol. 1, pp. 283–293). Springer.

- [24] Patro, C., & Pan, I. (2021). Decision tree-based classification model to predict student employability. In *Research and Applications in Artificial Intelligence (RAAI 2020)* (pp. 220–230). Springer.
- [25] Sobnath, D., et al. (2020). Feature selection for UK disabled students' engagement post higher education: A machine learning approach for a predictive employment model. *IEEE Access*, 8, 159530–159541.
- [26] Bai, A., & Hira, S. (2021). An intelligent hybrid deep belief network model for predicting students' employability. *Soft Computing*, 25(14), 9241–9254.
- [27] Ambili, B. A. (2024). A comprehensive evaluation of employability prediction using ensemble learning techniques. *EPRA International Journal of Multidisciplinary Research*, 10(1), 362–366.
- [28] Kahlik, A. A. B., & Saleh, A. Y. (2024). Predicting post-internship employability using ensemble machine learning approach. *Journal of Cognitive Sciences and Human Development*, 10(2), 87–101.
- [29] Alzahrani, M. R. (2024). Predicting student performance using ensemble models and learning analytics techniques. Preprints. <https://doi.org/10.20944/preprints202406.1100.v1>
- [30] Ali, M. F., Elamparithi, M., & Anuratha, V. (2024). Students' academic performance prediction using an ensemble of machine learning models. *Journal of Computer Analysis and Applications*, 33(4), 110–116.
- [31] Pandya, A., et al. (2024). Predictive modeling of engineering placement outcomes using genetic algorithms. In *Proceedings of the International Conference on Signal Processing and Advanced Research in Computing (SPARC)* (Vol. 1, pp. 45–52). IEEE.
- [32] Muhammad, L. A., & Argungu, M. S. (2023). Ensemble approach for predicting students' academic performance. *International Journal of Education and Pedagogical Sciences*, 17(2), 111–117.
- [33] Vo, M. T., Nguyen, T., & Le, T. (2023). OPT-BAG model for predicting student employability. *Computers, Materials & Continua*, 76(2), 240–250.
- [34] Sun, T., & He, Z. (2023). Developing intelligent hybrid DNN model for predicting students' employability: A machine learning approach. *Journal of Education, Humanities and Social Sciences*, 18, 235–248.
- [35] Darji, J. H., Nakrani, T. V., Sandhi, I. I., Patel, R. D., Patel, P. A., & Raval, P. D. (2022). Predicting students' grades in a professional undergraduate course using an ensemble model. *Mathematics, Statistics & Engineering Applications*, 71(4), 9356–9370.
- [36] Krishnaiah, V., & Kadegowda, Y. H. (2022). Undergraduate engineering students' employment prediction using hybrid approach in machine learning. *International Journal of Electrical and Computer Engineering*, 12(3), 2783–2792.
- [37] Maaliw, R. R., Quing, K. A. C., Lagman, A. C., Ugalde, B. H., Ballera, M. A., & Ligayo, M. A. D. (2022). Employability prediction of engineering students using ensemble classification modeling. In *Proceedings of the IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 288–294).
- [38] Saidani, O., Menzli, L. J., Ksibi, A., Alturki, N., & Alluhaidan, A. S. (2022). Predicting student employability through the internship context using gradient boosting models. *IEEE Access*, 10, 46472–46489.
- [39] Pandey, A., & Maurya, L. S. (2021). Prediction of undergraduate students' career using various machine learning and ensemble learning algorithms. *Webology*, 18(6), 23–35.
- [40] He, S., Li, X., & Chen, J. (2021). Application of data mining in predicting college students' employment. In *Proceedings of the 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)* (pp. 65–69).
- [41] Kumar, D., Verma, C., Singh, P. K., Raboaca, M. S., Felseghi, R. A., & Ghafoor, K. Z. (2021). Computational statistics and machine learning techniques for effective decision making on students' employment for real-time. *Mathematics*, 9(11), 1166.
- [42] Premalatha, N., & Sujatha, S. (2021). An effective ensemble model to predict employment status of students in higher educational institutions. In *Proceedings of the 4th International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (pp. 1–4).
- [43] Musiliu, B. (2020). Single classifiers and ensemble approach for predicting students' academic performance. *International Journal of Research and Scientific Innovation*, 7(6), 238–243.
- [44] Yan, L., & Liu, Y. (2020). An ensemble prediction model for potential student recommendation using machine learning. *Symmetry*, 12(5), 728.
- [45] Dutta, S., & Bandyopadhyay, S. K. (2020). Forecasting of campus placement for students using ensemble voting classifier. *Asian Journal of Research in Computer Science*, 5(4), 1–12.
- [46] Farissi, A., Dahlan, H. M., & Samsuryadi. (2020). Genetic algorithm-based feature selection for predicting students' academic performance. In *Emerging Trends in Intelligent Computing and Informatics* (Vol. 4, pp. 155–164). Springer.
- [47] Zhao, Zhiqiang, and Ping Ren. "Random Forest-Based Early Warning System for Student Dropout Using Behavioral Data." *Bulletin of Education and Psychology* (2025).

- [48] Nayak, Janmenjoy, Bighnaraj Naik, and H. S. Behera. "A comprehensive survey on support vector machine in data mining tasks: applications & challenges." *International Journal of Database Theory and Application* 8.1 (2015): 169-186.
- [49] Wang, Lishan. "Research and implementation of machine learning classifier based on KNN." *IOP Conference series: materials science and engineering*. Vol. 677. No. 5. IOP publishing, 2019.
- [50] A. Moutni, Students' Employability Dataset, Kaggle, 2023. [Online]. Available: <https://www.kaggle.com/datasets/anashamoutni/students-employability-datase>