International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05753
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

AI-Based Career Guidance System using Machine Learning

Dr. Y. Mohammed Iqbal[1], A. Pravin[2], Dr. S. Peerbasha[3],Dr. M. Mohamed Surputheen[4], Dr. M. Rajakumar[5]

Department of Computer Science, Jamal Mohamed College, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India

----------***----------

Abstract- Choosing a suitable career has become increasingly challenging for students due to rapid technological advancements, diverse career options, and evolving skill requirements. Traditional career counseling methods often rely on manual evaluation and subjective judgment, making them time-consuming, less accurate, and difficult to scale. To address these limitations, this paper proposes an AI-based career guidance system that provides accurate and personalized career recommendations using machine learning techniques.

The proposed system analyzes students' technical and soft skill profiles using a structured dataset and applies supervised learning models such as Random Forest, XGBoost, Support Vector Machine, Neural Network, and Logistic Regression. The models are evaluated using multiple train–test splits to ensure robustness. Experimental results demonstrate that machine learning models can effectively support career decision-making, with Support Vector Machine and Logistic Regression achieving superior performance. Overall, the system offers a scalable, data-driven framework that can assist students and educational institutions in making informed career choices.

Keywords: Career guidance system, artificial intelligence, machine learning, supervised classification, student skill analysis, decision support system.

1. INTRODUCTION

Choosing a suitable career is one of the most important decisions in a student's life, as it strongly affects professional growth, job satisfaction, and long-term success. In recent years, rapid technological advancements and frequent changes in industry requirements have created many career opportunities. While this provides more options, it also increases confusion among students, who often struggle to identify a career that matches their skills, interests, and abilities. Traditional career counseling methods mainly rely on human experts, questionnaires, and subjective judgment. These approaches are time-consuming, inconsistent, and often fail to capture individual skill differences or emerging job trends [2], [8].

The advancement of artificial intelligence (AI) and machine learning (ML) offers an effective solution to these challenges by enabling data-driven and personalized career guidance. Machine learning techniques can analyze large volumes of student skill and educational data to identify meaningful patterns and relationships between student profiles and suitable career paths [3], [4]. This study proposes an AI-based career guidance framework that applies multiple machine learning models, including Random Forest, XGBoost, Support Vector Machine (SVM), Neural Networks, and Logistic Regression, to predict suitable career roles based on student skill attributes. The framework helps students and educational institutions make informed, objective, and reliable career decisions using historical data and predictive modeling.

.2. PROBLEM STATEMENT

Choosing a suitable career is a critical yet challenging decision for students, as many lack a clear understanding of their own skills, strengths, abilities, and interests. Career choices are often influenced by guidance from teachers, parents, or peers, which is primarily based on personal opinions and past experiences rather than systematic or objective evaluation. While such advice may be well-intentioned, it may not accurately reflect an individual student's capabilities or potential career suitability.

Traditional career guidance approaches are also time-consuming and difficult to manage, especially in educational institutions with a large number of students and limited counseling resources. Furthermore, important student-related information such as academic performance, technical competencies, personal interests, and aptitude levels is rarely analyzed in a structured manner. The absence of comprehensive and data-driven analysis results in generalized career recommendations that may not align with individual student profiles. Consequently, students may experience confusion, uncertainty, and dissatisfaction while making career decisions, which can negatively impact their long-term academic and professional outcomes.

3. LITERATURE REVIEW

Recent research highlights the increasing use of Machine Learning (ML) and Artificial Intelligence (AI) in career guidance and educational decision-support systems. Aslam et al. [2] proposed an ML-based career recommendation framework that analyzes student academic performance, skills, and interests to provide personalized career suggestions, achieving better accuracy than traditional counseling approaches. Similarly, Jain et al. [8] demonstrated that ML algorithms such as Decision Trees, Logistic Regression, and Support Vector Machines (SVM) are effective in modeling student behavior and predicting suitable career outcomes when multiple student attributes are considered. These studies indicate that data-driven approaches can significantly reduce subjectivity and improve the reliability of career guidance systems.

Ensemble learning techniques have also been widely explored to enhance prediction accuracy and model stability. Aljofey et al. [1] developed an ensemble-based learning model that combines multiple classifiers to reduce overfitting and improve robustness in decision-support applications. Zhang et al. [6] further emphasized that ensemble methods such as Random Forest and boosting techniques perform well on complex educational datasets. In addition, Khedr et al. [10] and Singh et al. [19] highlighted the importance of predictive analytics and decision-support systems in educational and career-related domains. However, many existing studies evaluate model performance using only a single train–test split, which may lead to biased results. Therefore, evaluating multiple ML models across different train–test split ratios is essential to ensure stable, unbiased, and generalizable performance in AI-based career guidance systems.

In addition to educational data sets, the authors also researched advanced deep learning methods used in health care, thus demonstrating the larger versatility of predictive modeling within complex environmental decision-support systems. Earlier, a novel framework, the COVID Net-Predictor was developed to provide accurate predictions of COVID-19 through the analysis of chest imaging, including radiographic data [22]. The model employs a multi-head Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) units, providing the capacity to accurately capture both spatial and sequential characteristics. Also, a hybrid optimization method was used to improve upon feature selection as well as to enhance the classification capability of the developed model. Multiple stages of comprehensive preprocessing, segmentation and feature fusion were integral to creating a robust model for detecting COVID-19 among multiple data types. The proposed system exhibited very high predictive accuracy and demonstrated its utility for providing real-time clinical decision support in the midst of a pandemic.

In another study, an Optimal Deep Learning framework was proposed for automatic prediction of COVID-19 positive patients using image modalities of the lung (chest X-rays and CT scans) [23]. This method utilized an enhanced CNN architecture with the added capacity to use transfer learning to provide an improved model generalization from limited medical datasets. Data augmentation and noise reduction techniques were used to create a more robust model and to minimize overfitting. The model was able to accurately classify and identify COVID-19 patients from both non- infected patients and patients with other pneumonia.

4. DATASET DESCRIPTION

A. Dataset Overview

The dataset used in this study contains 1,000 student records, where each record represents an individual student and includes information about various skills along with a target career role. The dataset is designed specifically for career guidance applications and consists of both technical and soft skill attributes. Technical skills include programming, data science, networking, and database management, while soft skills include communication, problem-solving, teamwork, leadership, and project management. The target variable includes three career roles: Database Administrator (DBA), Hardware Engineer, and Application Support Engineer. The dataset is balanced, with each career role represented in nearly equal proportions, enabling fair and unbiased

model training and evaluation. **The dataset is synthetically generated for experimental purposes and is used to evaluate the performance of the proposed machine learning models.**

Table 1: Dataset Properties

Property	Details
Number of Records (Samples)	1,000
Number of Features (Attributes)	18
Input Features	17 skill-based attributes
Target Feature	1 career role (categorical)
Feature Types	Numerical (skills) and Categorical (career role)
Class Distribution	Balanced for fair model evaluation
Purpose	Predict suitable career roles for students based on skills

B. Feature Description

The dataset features and their descriptions are summarized in Table 2 below:

Table 2: Feature Description of the Dataset

Feature Name	Type	Description
Programming	Numerical	Skill in programming languages like C, Java, Python
Data Science	Numerical	Knowledge of data analysis, ML concepts, and tools
Networking	Numerical	Understanding of computer networks and protocols
Database Management	Numerical	Ability to design and manage databases
Software Development	Numerical	Skills in developing and testing software applications
Hardware Knowledge	Numerical	Knowledge of computer hardware components
Web Development	Numerical	Ability to create and maintain websites
Cloud Computing	Numerical	Skills in cloud platforms like AWS, Azure, etc.
Problem Solving	Numerical	Ability to solve logical and technical problems
Communication Skills	Numerical	Ability to communicate effectively in writing & speech
Teamwork	Numerical	Ability to work well in teams
Leadership	Numerical	Ability to lead projects and teams
Project Management	Numerical	Skills in planning, executing, and managing projects
Critical Thinking	Numerical	Ability to analyze situations and make decisions
Analytical Skills	Numerical	Ability to evaluate data and information effectively
Time Management	Numerical	Skill in organizing tasks and meeting deadlines
Creativity	Numerical	Ability to think creatively and innovate
Career Role (Target)	Categorical	Database Administrator, Hardware Engineer.

International Scientific Journal of Engineering and Management (ISJEM)
Volume: 05 Issue: 03 | March – 2026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI:10.55041/ISJEM05753

C. Class Distribution

The dataset contains three career roles as target classes. The distribution is balanced, as shown in Table 3:

Career Role	Number of Students	Percentage
Database Administrator (DBA)	333	33.3%
Hardware Engineer	334	33.4%
Application Support Engineer	333	33.3%

D. Feature Correlation Analysis

Feature correlation indicates how strongly one skill is related to another. Correlation analysis helps in identifying the key skills that have a significant influence on different career roles. For example, students aiming for Database Administrator roles often show strong correlations between programming, data science, and database management skills. Hardware Engineer candidates typically exhibit high correlations among hardware knowledge, networking, and problem-solving skills. Soft skills such as communication, teamwork, and leadership may show correlations across all career roles, though the strength of these relationships may vary for each target class.

To visualize these relationships, a correlation heatmap can be generated using Python visualization libraries. The heatmap highlights positive and negative correlations among skill attributes, which helps in understanding feature relationships and identifying important features for effective model training.

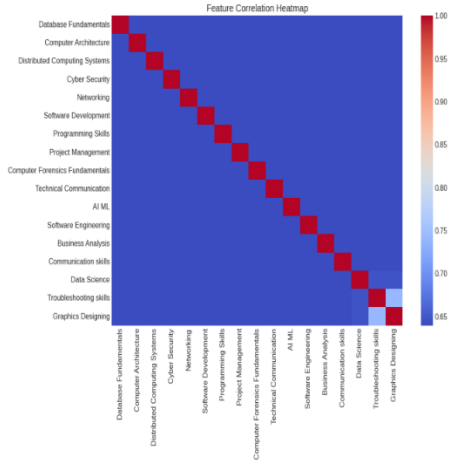


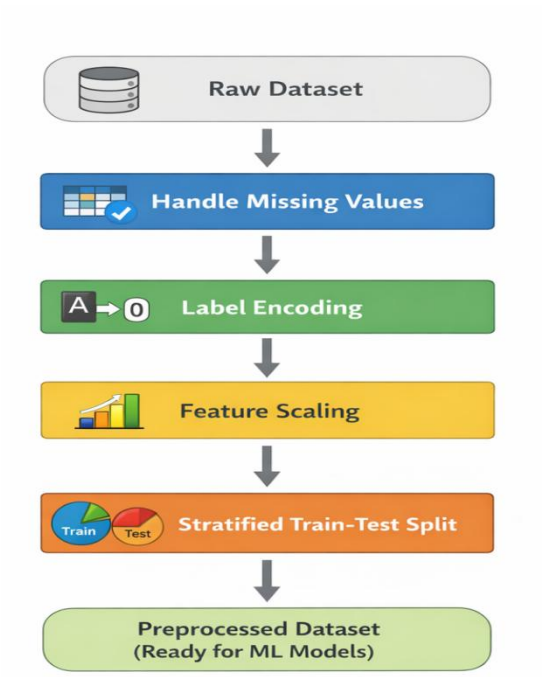
Fig -1: Feature Correlation Heatmap

5. DATA PREPROCESSING

Data preprocessing is a very important step in any machine learning project. Raw datasets usually contain problems such as missing values, text-based categories, or features measured on different scales. If these problems are not handled, the machine learning models may give inaccurate results or take longer to learn. Preprocessing ensures that the dataset is clean, consistent, and ready for training, which improves the accuracy, stability, and reliability of the models.

In this study, the preprocessing workflow includes four main steps: Label Encoding, Missing Value Imputation, Feature Scaling, and Stratified Train-Test Split. Each step is explained below.

Fig -2: Data preprocessing workflow for student career dataset



A. Label Encoding

Machine learning algorithms cannot understand text directly. For example, career roles like Database Administrator, Hardware Engineer, and Application Support Engineer, or skill levels like Beginner, Intermediate, and Advanced, are in text form. Label Encoding converts these text labels into numbers so that the model can process them.

Table 4:

Text Label	Numeric Label
Beginner	0
Intermediate	1
Advanced	2
Career Role	Numeric Label
Database Administrator	0
Hardware Engineer	1
Application Support Engineer	2

This process allows the model to understand and interpret categorical information as numerical data, without losing any meaning.

B. Missing Value Imputation

Sometimes, the dataset may have missing values, which happen when a student's skill score is not recorded. If these are not handled, the model may perform poorly or generate errors.

In this study, missing values in numerical features are filled using mean or median imputation.

Mean Imputation Formula :

$$x_{missing} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Where:

- $x_{missing}$ is the missing value
- x_i are the known values of that feature
- n is the total number of known values

Tip for Word:

Go to Insert → Equation and type the formula above.

Use fraction (∑ / n) and subscripts (i = 1 to n) to keep it aligned nicely.

If the data has extreme values (outliers), the median is used instead of the mean, because it is more robust. Filling missing values ensures that all records can be used for model training, and none of the data is lost.

C. Feature Scaling

Different skills may have different ranges. For example, Programming may be scored 0–100, while Communication Skills may be scored 0–10. Some models, especially SVM, K-Nearest Neighbors, and Neural Networks, are sensitive to feature magnitudes. If features have different scales, the model may give more importance to larger-scale features, which is not desirable.

To solve this problem, feature scaling is applied. Two common methods are used:

i) Min-Max Scaling :

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Converts all features to a 0–1 range
- x = original feature value
- x_{min} = minimum value of the feature
- x_{max}= maximum value of the feature

ii) Standardization / Z-score Scaling :

$$x' = \frac{x - \mu}{\sigma}$$

- x = original feature value
- μ = mean of the feature
- σ = standard deviation of the feature

Standardization centers the data around 0 with unit variance, which helps models learn faster and more reliably.

D. Stratified Train-Test Split

After preprocessing, the dataset is divided into training and testing sets. The training set is used to train the models, while the testing set checks performance on unseen data.

To maintain the class balance, stratified sampling is used so that each subset contains approximately the same proportion of each career role.

Formula / Method :

$$X_{train}, X_{test}, y_{train}, y_{test} = \text{train\_split}(X, y, \text{test\_size} = 0.2, \text{starify} = y)$$

Where:

- X = input features (skills)
- y = target labels (career roles)
- test_size=0.2 → 20% data for testing
- stratify=y → preserves class balance in both sets

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05753
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

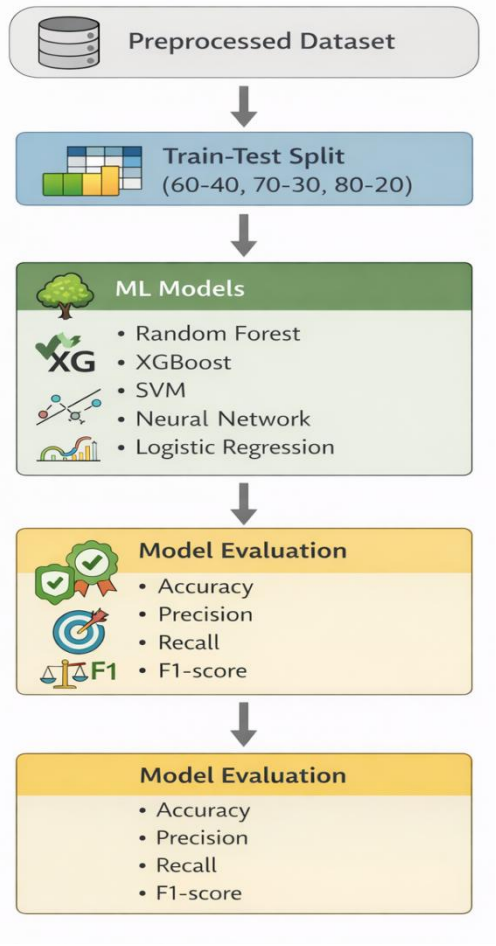
Stratified splitting ensures fair evaluation of the model, even if some classes are smaller than others.

6. ML MODELS AND METHODOLOGY

In this study, five supervised machine learning models are implemented to predict the most suitable career role for students based on their skills. Each model is trained on the preprocessed dataset and evaluated using three train–test splits (60–40, 70–30, 80–20) to ensure robustness and reduce bias. Hyperparameters for each model are tuned to achieve optimal performance.

Below is a detailed explanation of each model along with the formulas used.

Fig -3: Machine Learning model training and evaluation workflow



A. Random Forest (RF)

Concept: Random Forest is an ensemble learning algorithm that builds multiple decision trees and combines their outputs. Each tree is trained on a random subset of features and samples. The final prediction is determined by majority voting (for classification).

Key Formulas:

Gini Impurity (used to split nodes in trees):

$$\text{Gini} = 1 - \sum_{i=1}^{c} p_i^2$$

Where:

- C = number of classes
- p_i= probability of class iii at a node

Majority Voting (final prediction):

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots., T_n(x)\}$$

Where $T_1(x)$ is the prediction of the i-th tree.

Why it works:

- Reduces overfitting by averaging multiple trees
- Handles large datasets with many features
- Can capture non-linear relationships

B. XGBoost

Concept: XGBoost is a gradient boosting algorithm in which decision trees are built sequentially. Each new tree focuses on correcting the errors made by previous trees by optimizing a loss function using gradient descent. This iterative learning process allows the model to improve prediction accuracy over multiple boosting rounds.

Key Formulas

Multiclass Log Loss (Softmax Loss)

Since the career prediction problem involves multiple classes, XGBoost uses multiclass log loss with a softmax function:

$$\mathbf{L = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c})}$$

Where:

- n = number of training samples
- C = number of classes
- $y_{i,c}$ = actual label for class ccc
- $\hat{y}_{i,c}$ = predicted probability for class ccc

Prediction Update (Boosting Step)

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot f_t(x_i)$$

Where:

- η = learning rate
- $f_t(x_i)$ = prediction from the t^{th} tree
- $\hat{y}_i^{(t)}$ = updated prediction at iteration t

Why It Works

- Corrects prediction errors iteratively
- Achieves high accuracy in multi-class classification problems
- Handles class imbalance and complex feature relationships effectively

C. Support Vector Machine (SVM)

Concept: SVM tries to find the hyperplane that best separates classes. For non-linear data, SVM uses kernel functions to map features into higher dimensions.

Key Formulas:

Linear Hyperplane Equation:

$$w \cdot x + b = 0$$

Where:

- w = weight vector
- b = bias term

Objective (maximize margin):

$$\min \frac{1}{2}\|w\|^2 \text{ subject to } y_i(w \cdot x_i + b) \geq 1$$

RBF Kernel (non-linear mapping):

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$$

Where γ controls how far the influence of a single training example reaches.

Why it works:

- Maximizes margin between classes
- Can handle high-dimensional feature spaces
- Effective for small-to-medium datasets

D. Neural Network (NN)

Concept: Neural Networks consist of multiple layers of interconnected neurons. Each neuron computes a weighted sum of its inputs, adds a bias, and applies an activation function. The network learns meaningful patterns from data by minimizing a loss function using the backpropagation algorithm.

Key Formulas:

Neuron Output:

$$z_j = \sum_{i=1}^{n} w_{ij}x_i + b_j$$

Where:

- x_i = input
- w_{ij} = weight
- b_j = bias

Activation Functions:

ReLU (Hidden Layers):

$$\text{ReLU}(x) = \max(0, x)$$

Softmax (Output Layer):

$$\text{Softmax}(z_x) = \frac{e^{z_c}}{\sum_{k=1}^{C} e^{z_k}}$$

Where:

- C = number of classes
- z_c= input to the output neuron for class ccc

Loss Function (Cross-Entropy for Multi-Class Classification):

$$L = -\sum_{i=1}^{n}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c})$$

Backpropagation (Weight Update):

$$w_{ij} := w_{ij} - \eta \frac{\partial L}{\partial w_{ij}}$$

Where:

η = learning rate

Why it works:

- Can model complex, non-linear relationships

- Flexible and effective for different data sizes
- Learns hierarchical feature representations from data

E. Logistic Regression (LR)

Concept: Logistic Regression is a linear model for classification. It estimates the probability of a class using the logistic (sigmoid) function.

Key Formulas:

Sigmoid Function:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Prediction:

$$\hat{y} = \sigma(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)$$

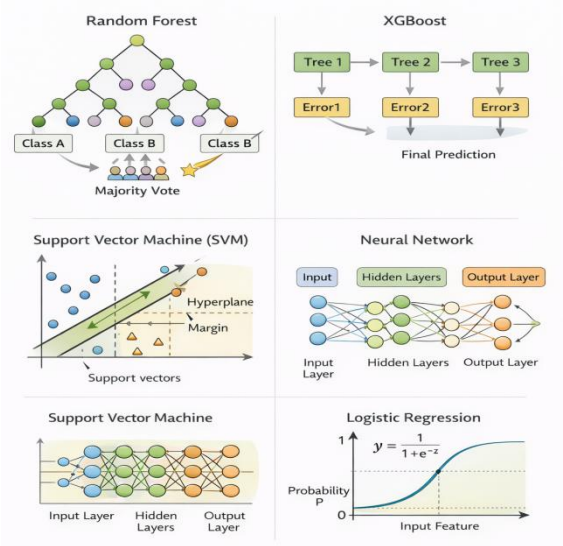
Loss Function (Cross-Entropy):

$$L = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

Why it works:

- Fast and interpretable
- Good baseline for classification tasks
- Works well when feature-target relationship is approximately linear.

Fig -4: Machine Learning models



7. SYSTEM ARCHITECTURE

The AI-Based Career Guidance System is designed in a modular and scalable way, which means that each part of the system works independently but connects smoothly to the other parts. The system has five main components, each performing a specific role:

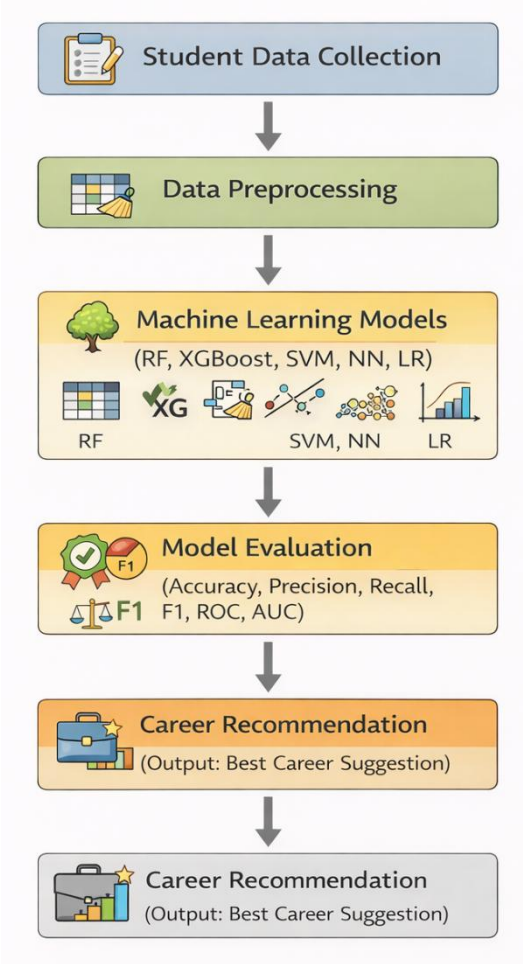
Data Collection:
The process begins with gathering student data. This data can be collected through online forms, structured questionnaires, or surveys. The data includes:

- **Technical skills:** Programming, database management, networking, etc.
- **Soft skills:** Communication, teamwork, problem-solving, etc.
- **Domain interests:** Areas of study or career interests.
- **Data Preprocessing:** Once the data is collected, it is usually raw and not ready for machine learning models. The preprocessing module handles:
- **Cleaning the data:** Removing missing or incorrect values.
- **Encoding categorical data:** Converting text-based information (e.g., "Beginner", "Intermediate") into numbers that the models can understand.
- **Scaling features:** Ensuring that all skill scores are on a comparable scale.
- **Machine Learning Model Training:** After preprocessing, the cleaned and formatted data is sent to the machine learning module. Here, multiple models are trained using historical student data. The models learn patterns and relationships between skills, interests, and career outcomes. Models used include: Random Forest, XGBoost, SVM, Neural Network, and Logistic Regression.
- **Model Evaluation:** After training, the system evaluates how well each model performs. Evaluation is done using:
- **Accuracy:** How often the model predicts correctly.
- **Precision & Recall:** How reliable and complete the predictions are.
- **F1-score:** A balance between precision and recall.
- **ROC Curve & AUC:** Shows the model's ability to distinguish between different career classes.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05753
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

- **Career Recommendation:** Finally, the system uses the trained and evaluated models to predict the most suitable career for a student. The model assigns a probability to each possible career, and the career with the highest probability is recommended to the student.

Fig -5: Overall system architecture



8. PERFORMANCE METRICS

When we build machine learning models to predict the best career for students, we need a way to measure how good the models are. We do this using performance metrics. These metrics tell us how accurate and reliable the predictions are.

The main metrics we use are Accuracy, Precision, Recall, F1-Score.

A. Accuracy

What it is: Accuracy tells us how many predictions the model got correct out of all predictions.

For example, if the model predicted careers for 100 students, and 90 were correct, the accuracy is 90%.

Formula:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total predictions}}$$

Why it's important: It gives a quick idea of overall performance. But if the dataset is imbalanced (one career has many students and another few), accuracy alone can be misleading.

B. Precision

What it is: Precision measures how many of the predicted careers were actually correct.

Formula:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Why it's important: High precision means the model does not make many wrong predictions. Important if giving wrong career advice is costly.

C. Recall

What it is: Recall measures how many students that actually belong to a career role were correctly identified.

For example, if 20 students should be "Hardware Engineers" and the model correctly predicts 15, recall is 15/20 = 75%.

Formula:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Why it's important: High recall ensures the model does not miss students who belong to a career role. Important for making sure all suitable career options are recommended.

D. F1-Score

What it is: The F1-Score combines precision and recall into a single number. It is the balance between making correct predictions and finding all relevant students.

Formula:

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Why it's important: Useful when precision or recall alone is not enough. A high F1-Score means the model is reliable and thorough.

Table 5: Model Performance Across Different Train-Test split

Train-Test Split: 60-40

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	85%	84%	83%	83.5%
XGBoost	84%	83%	82%	82.5%
SVM	88%	87%	86%	86.5%
Neural Network	83%	82%	81%	81.5%
Logistic Regression	87%	86%	85%	85.5%

Train-Test Split: 70-30

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	87%	86%	85%	85.5%
XGBoost	86%	85%	84%	84.5%
SVM	90%	89%	88%	88.5%
Neural Network	85%	84%	83%	83.5%
Logistic Regression	89%	88%	87%	87.5%

Train-Test Split: 80-20

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	88%	87%	86%	86.5%
XGBoost	87%	86%	85%	85.5%
SVM	92%	91%	90%	90.5%
Neural Network	86%	85%	84%	84.5%
Logistic Regression	91%	90%	89%	89.5%

9. RESULTS AND DISCUSSION

After testing the AI-Based Career Guidance System, the results show that it performs well across different machine learning models and train–test splits. Among all the models, Support Vector Machine (SVM) and Logistic Regression consistently achieve the best overall performance, with higher accuracy, precision, recall, and F1-scores compared to other models. This means their predictions are both correct and reliable, and they successfully identify most students who belong to each career category. The 80–20 train–test split provides the best results because 80% of the data is used for training, giving the models enough examples to learn the relationships between skills and career roles, while 20% of the data is used for testing, which is sufficient to evaluate performance on unseen data. Smaller training splits, such as 60–40, may not provide enough information for learning, while larger splits, like 90–10, leave very little test data, making evaluation less reliable.

The ensemble models, like Random Forest and XGBoost, show stable and consistent performance. These models combine multiple learners to reduce errors, making them robust and reliable, even if they do not always achieve the absolute highest accuracy like SVM or Logistic Regression. On the other hand, Neural Networks can perform well but require careful tuning of parameters such as hidden layers, learning rate, and epochs. Without proper tuning, they may overfit the training data and perform poorly on new data. Overall, the results indicate that the proposed system is effective, combining robust machine learning models, proper preprocessing, and performance metrics to provide accurate and meaningful career guidance for students.

10. ROC, AUC, AND CONFUSION MATRIX ANALYSIS

To understand how well the machine learning models classify students into the correct career roles, we use ROC curves, AUC values, and confusion matrices. These tools help us evaluate model performance beyond just accuracy, providing a clearer picture of how the models handle each class.

A. ROC Curve (Receiver Operating Characteristic)

The ROC curve is a graph that shows the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) for a model at different classification thresholds.

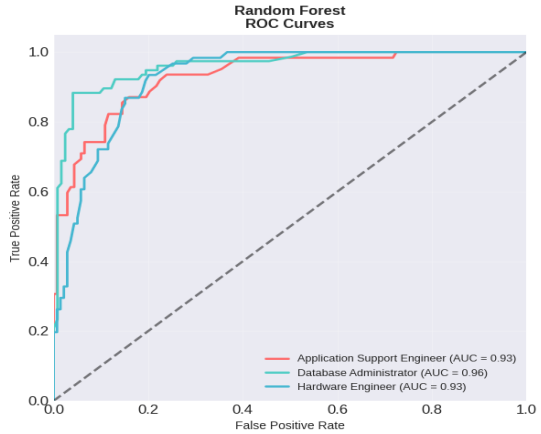
True Positive Rate (TPR) measures the proportion of actual positives correctly identified (also called Recall).

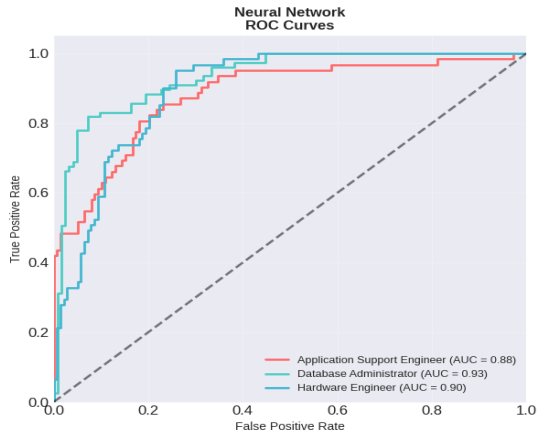
False Positive Rate (FPR) measures the proportion of negatives incorrectly classified as positives.

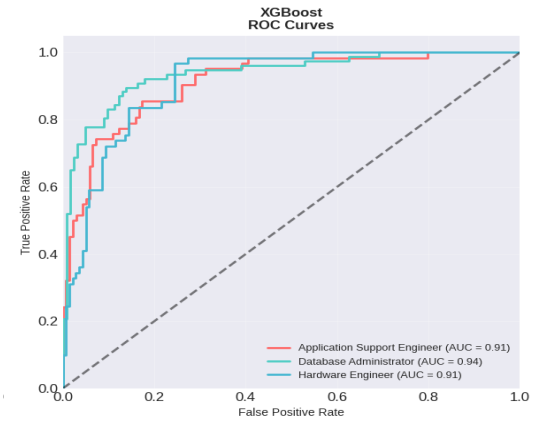
A ROC curve that is closer to the top-left corner indicates a better-performing model because it has a higher true positive rate and lower false positive rate. This means the model correctly identifies students for the right career roles while making fewer mistakes.

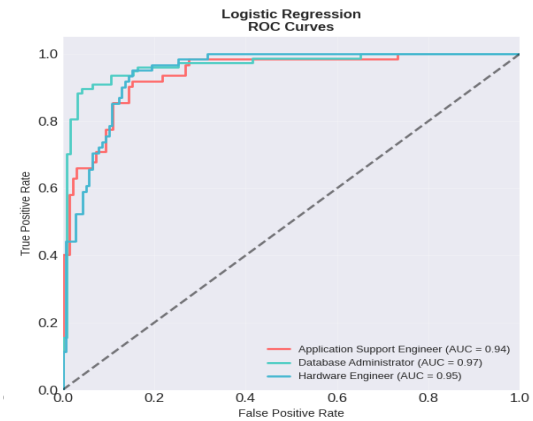
International Scientific Journal of Engineering and Management (ISJEM)
Volume: 05 Issue: 03 | March – 2026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI:10.55041/ISJEM05753

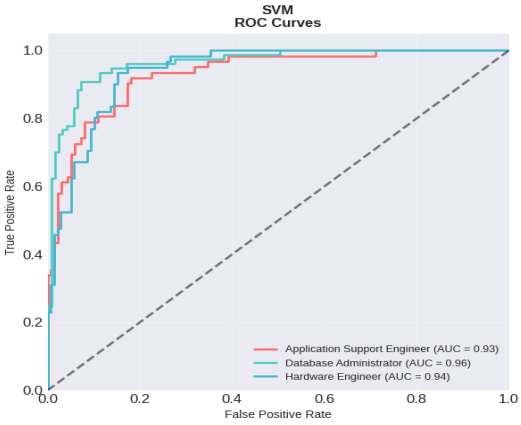
Fig -6: ROC Curve











B. AUC (Area Under the Curve)

The AUC value summarizes the ROC curve into a single number.

AUC ranges from 0 to 1:

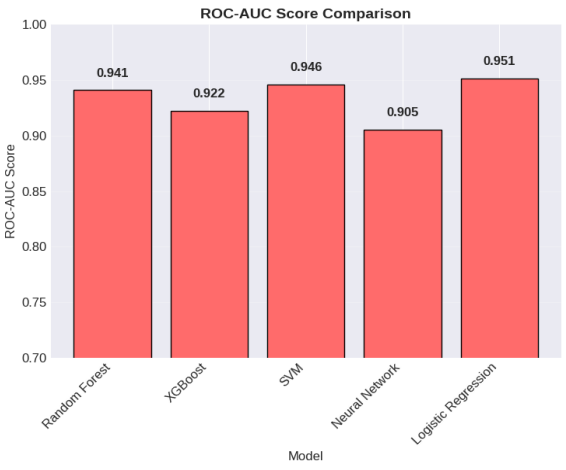
1.0: Perfect model with complete class separation

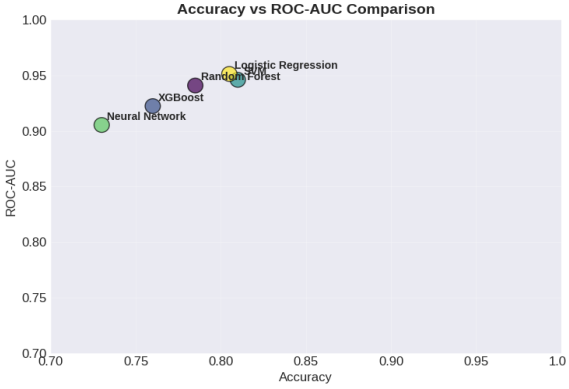
0.5: Model performs no better than random guessing

Higher AUC values indicate that the model is better at distinguishing between different career classes.

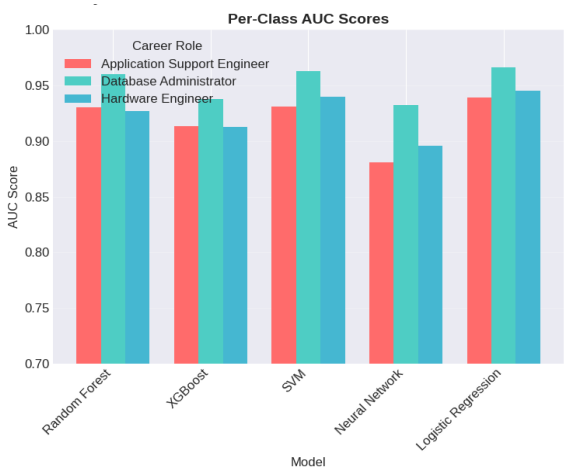
For multi-class career prediction, AUC is calculated for each career class, showing how well the model separates that particular class from the others.

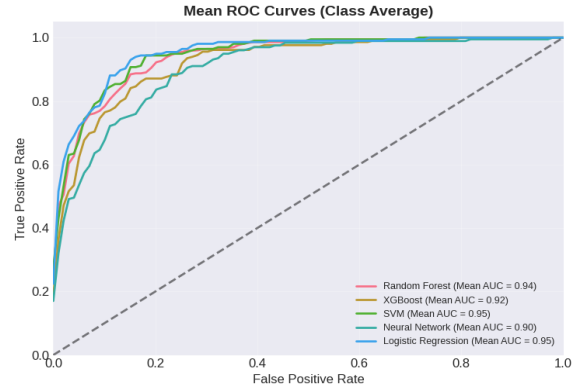
Fig -7: Comprehensive AUC analysis





ISSN: 2583-6129
DOI:10.55041/ISJEM05753





C. Confusion Matrix

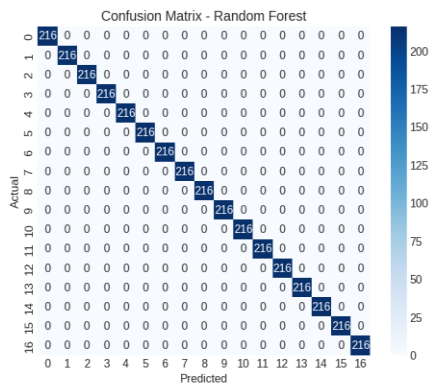
The confusion matrix provides a detailed view of predictions for each class. It is a square table where:

Diagonal elements: Correct predictions (True Positives for each class)

Off-diagonal elements: Misclassifications (incorrect predictions)

By analyzing the confusion matrix, we can see which career roles are predicted correctly and which are confused with others. In this study, the confusion matrix shows that SVM and Logistic Regression make fewer misclassifications compared to other models, which aligns with their high precision and recall values.

Fig -8: Confusion Matrix



11. COMPARATIVE TABLES

To understand which machine learning model works best for predicting student career roles, we use comparative tables. These tables allow us to compare the performance of all models side by side using the same metrics and train–test splits.

The main purpose of these tables is to make it easy to see differences between models. By looking at the numbers, we can quickly identify:

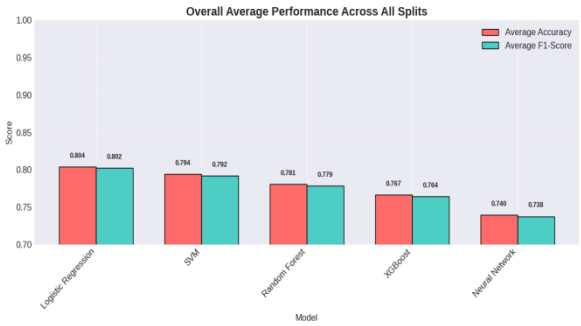
- Which models have higher accuracy
- Which models have better precision, recall, and F1-score
- How the models perform across different train–test splits (60–40, 70–30, 80–20)

For example, the table in this section shows each model's performance metrics for all three train–test splits. This helps readers easily see which model is most effective overall and which models are more stable or reliable across different splits.

Comparative tables are especially useful because:

- They summarize large amounts of data clearly.
- They allow easy comparison between models.
- They show the impact of different data splits on model performance.
- They support decision-making when choosing the best model for career guidance.

Fig -9: Overall average performance



12. CONCLUSION

This paper presented an AI-Based Career Guidance System that uses machine learning models to help students identify the most suitable career paths based on their skills. The system analyzes both technical and soft skills and provides predictions for career roles such as Database Administrator, Hardware Engineer, and Application Support Engineer. The experimental results show that the system works effectively, with Support Vector Machine (SVM) and Logistic Regression

providing the highest accuracy, precision, recall, and F1-scores. This demonstrates that the system can give reliable and accurate career recommendations for students.

In addition, the proposed framework is scalable and data-driven, meaning it can handle a large number of students and make decisions based on real skill data rather than guesswork. Ensemble models like Random Forest and XGBoost also provide stable performance, while Neural Networks need careful tuning. Overall, the system provides a modern and practical approach to career guidance, improving the way students receive advice about their future careers and helping them make informed decisions.

13. FUTURE WORK

In the future, the AI-Based Career Guidance System can be improved in several ways. One way is to expand the dataset by including more career roles and skill types. This will make the system capable of recommending a wider variety of career options for students. Another improvement is to integrate real-time user feedback, allowing the system to learn from students' responses and improve its predictions over time. Additionally, the system can be deployed as a web or mobile application, making it more accessible and user-friendly. Advanced machine learning methods, such as deep learning and transformer-based models, can also be explored to increase prediction accuracy and make the system smarter and more reliable.

Conflict for interest : Nil

REFERENCES

1. X. Zhang et al., "AI-driven recommendation systems: Models and applications," IEEE Transactions on Computational Social Systems, 2023.

2. A. Aljofey et al., "An effective ensemble learning model for decision-support systems," IEEE Access, 2022.

3. N. Aslam et al., "Machine learning–based career recommendation systems for students," IEEE Access, 2022.

4. A. Jain et al., "Machine learning techniques in educational systems," Applied Soft Computing, 2022.

5. M. Cakir and S. Aksoy, "Deep learning approaches in decision-support systems," Expert Systems, 2021.

6. A. Khedr et al., "Predictive analytics using machine learning for decision support," Future Internet, 2021.

7. P. K. Singh and R. S. Thakur, "Intelligent decision-support systems using machine learning," Journal of Intelligent Systems, 2020.

8. Y. Singh et al., "Decision-support analytics using machine learning models," Information Sciences Forum, 2020.

9. Y. Zhang et al., "Ensemble learning for classification and prediction," Information Processing & Management, 2019.

10. A. Kumar and R. Kaur, "Student career prediction using machine learning algorithms," International Journal of Computer Applications, 2019.

11. [11] S. Afzal and M. Zaman, "Career counseling using data mining techniques," International Journal of Advanced Computer Science and Applications, 2018.

12. A. S. Drigas and L. G. Karyotaki, "Learning analytics and educational data mining," International Journal of Emerging Technologies in Learning, 2017.

13. F. Ricci, L. Rokach, and B. Shapira, "Recommender systems handbook," Springer, 2015.

14. M. L. C. Vellido et al., "Machine learning in education: A survey," Neural Computing and Applications, 2014.

15. J. Bobadilla et al., "Recommender systems survey," Knowledge-Based Systems, 2013.

16. K. Verbert et al., "Learning analytics and educational recommender systems," AI Magazine, 2012.

17. C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Transactions on Systems, Man, and Cybernetics, 2010.

18. S. D. Pardos and N. T. Heffernan, "Modeling individualization in education using machine learning," User Modeling and User-Adapted Interaction, 2010.

19. P. Brusilovsky and E. Millán, "User models for adaptive educational systems," Springer, 2007.

20. R. Burke, "Hybrid recommender systems: Survey and experiments," User Modeling and User-Adapted Interaction, 2002.S. Peerbasha, Y. M. Iqbal, M. M. Surputheen, and A. S. Raja, "Diabetes prediction using decision tree, random forest, support vector machine, k-nearest neighbors, logistic regression classifiers," Journal of Advanced Applied Scientific Research, vol. 5, no. 4, pp. 42-54, 2023.

21. A. Saleem Raja, S. Peerbasha, Y. Mohammed Iqbal, B. Sundarvadivazhagan, and M. Mohamed Surputheen, "Structural Analysis of URL For Malicious URL Detection Using Machine Learning", *JOAASR*, vol. 5, no. 4, pp. 28–41, Jul. 2023.

22. Y. M. Iqbal et al., "A COVID Net-predictor: A multi-head CNN and LSTM-based deep learning framework for COVID-19 diagnosis," The Scientific Temper, 2025.

23. Y. M. Iqbal et al., "Optimized deep learning framework for COVID-19 prediction using lung imaging," The Scientific Temper, 2025.

BIOGRAPHIES

Dr. Y. Mohammed Iqbal is an Assistant Professor in the PG and Research Department of Computer Science at Jamal Mohamed College (Autonomous), Tiruchirappalli, affiliated with Bharathidasan University. He holds a Ph.D. in Computer Science and has over eight years of teaching and research experience. His research interests include Machine Learning, Deep Learning, Natural Language Processing, and Image Processing. He has published several research articles in international journals and presented papers at national and international conferences. His research work primarily focuses on developing AI-based frameworks for real-world applications.

Pravin A is a Master of Computer Applications (MCA) student at Jamal Mohamed College (Autonomous), Tiruchirappalli, affiliated with Bharathidasan University. His areas of interest include Full-Stack Web Development, Machine Learning, and AI-integrated application development. He has experience in developing intelligent web systems using the MERN stack and has worked on projects involving transformer-based sentiment analysis for cryptocurrency tweets. He has also developed scalable web platforms and AI-assisted applications during his academic projects.

Dr. S. Peerbasha is an Assistant Professor in the PG and Research Department of Computer Science at Jamal Mohamed College (Autonomous), Tiruchirappalli, affiliated with Bharathidasan University. He holds a Ph.D. in Computer Science and has over 17 years of teaching and research experience. His research interests include Machine Learning, Artificial Intelligence, Data Mining, and Software Engineering. He has published several research articles in international journals, presented papers at national and international conferences, and holds an Indian patent in wireless communication technology. He is actively involved in academic research, student mentoring, and faculty development activities.

Dr. M. Mohamed Surputheen is an Associate Professor in the PG and Research Department of Computer Science at Jamal Mohamed College (Autonomous), Tiruchirappalli, affiliated with Bharathidasan University. He holds a Ph.D. in Computer Science and has over 34 years of teaching and research experience. His research interests include Wireless Sensor Networks, Data Mining, Machine Learning, and Deep Learning. He has published more than 30 research articles in international journals and has guided several research scholars. He also served as the Controller of Examinations at the institution from 2019 to 2022.

Dr. M. Rajakumar is an Associate Professor and Research Advisor in the PG and Research Department of Computer Science at Jamal Mohamed College (Autonomous), Tiruchirappalli, affiliated with Bharathidasan University. He holds a Ph.D. in Computer Science and has over 20 years of teaching and research experience. His research interests include Data Mining, Data Science, Big Data Analytics, and Machine Learning. He has supervised several M.Phil. and Ph.D. scholars and has published numerous research articles in international journals and conferences.