Journal of Intelligent System and Applied Data Science (JISADS)

Journal homepage : https://www.jisads.com

ISSN (2974-9840) Online

Artificial Intelligence in Education Predicting College Plans of High School Students

Aws I. AbuEid 1,Wahida A. Mansouri[2], Achraf Ben Miled[2,3] , Ashraf F. A. Mahmoud[2], Faroug A. Abdalla[2], Chams Jabnoun[2], Aida Dhibi[2], Ahlem Fatnassi[2] , Firas M. Allan[2], Mohammed Ahmed Elhossiny[4,5], , Imen Ben Mohamed[2], Marwa Anwar Ibrahim Elghazawy[4], Majid A. Nawaz[2], Salem Belhaj[2]*

[1]*Faculty of Computing Studies, Arab Open University,Amman, Jordan*

[2]*Computer Science Department, Science College, Northern Border University, Arar, Kingdom of Saudi Arabia*

[3] *Artificial Intelligence and Data Engineering Laboratory, LR21ES23, Faculty of Sciences of Bizerte, University of Carthage, Tunisia*

[4]*Applied College, Northern Border University, Arar, Saudi Arabia*

[5]*Faculty of Specific Education, Mansoura University, Mansoura, Egypt.*

**Corresponding Author: Email: a_abueid@aou.edu.jo*

ABSTRACT

The study introduces AIRPCP (Artificial Intelligence for Educational Planning of College Pursuits), a predictive model utilizing artificial intelligence to forecast high school students' college plans. It achieves promising results with an accuracy rate of 84.88% through advanced techniques like hyperparameter tuning using scikit-learn's GridSearchCV. The model's development process, including data preprocessing and feature engineering, is detailed. Results show improved accuracy, precision, recall, and F1 scores, particularly for students with college plans. Visualizations aid in interpreting outcomes, supporting stakeholders in educational decision-making. The AIRPCP model has significant implications for educators, policymakers, and researchers, offering insights to inform educational planning and policy development, ultimately supporting students' aspirations.

Keywords: Artificial Intelligence, machine learning, python libraries, AIRPCP

1. INTRODUCTION

Pursuing a university education is a pivotal decision in the lives of high school students, significantly shaping their prospects and career paths. This critical choice carries immense weight for educational institutions and policymakers alike, as it forms the basis for planning admissions procedures and anticipating workforce requirements, thus informing management decision-making support strategies. Yet, obtaining precise insights into students' college aspirations presents a formidable challenge.

This article introduces a novel approach, the Artificial Intelligence in Education Predicting College Plans (AIEPCP) model, which leverages advanced artificial algorithms, particularly neural networks, to explore high school students' intentions regarding college attendance. These intentions are often considered highly personal and sensitive information, making students hesitant to share them openly [1]. This research aims to develop a new predictive model AIEPCP capable of estimating high school students' inclinations toward college

attendance. By doing so, it serves as a valuable resource for informing public universities and labor departments about the expected influx of new students or potential job seekers.

Furthermore, this study embarks on a journey into the domain of predictive analysis within the education sector, specifically focusing on forecasting high school students' college intentions. Employing the potent capabilities of Artificial Intelligence (AI), it harnesses the abundance of available student data. The ultimate objective is to discern the key factors differentiating high school students with a propensity for higher education from those who do not lean in that direction.

The motivation behind this research is rooted in the pressing need to address the challenges surrounding precise college enrollment predictions. These predictions are fundamental to educational planning and critical in anticipating labor market trends and optimizing resource allocation. Acknowledging the influential role of parents and the critical nature of factors such as gender, IQ, income, and parental encouragement [2], this study aspires to enhance decision-making processes within education.

Kaggle, a renowned platform for data science competitions, provided the dataset used in this study and datasets [3] sourced from government records. This dataset underwent rigorous curation and preprocessing to ensure data quality and privacy protection. Personal identifiers, such as student names, were thoughtfully desensitized to safeguard individual privacy and replaced with unique Student IDs.

The selected explanatory variables for analysis encompass gender, IQ, parental income, and parental encouragement to pursue higher education. According to experienced government officials, these variables have emerged as the most influential factors shaping high school students' college intentions [4]. Given the scale and intricacy of the dataset, this study offers access to a substantial sample consisting of up to 8,000 data samples. This sample size is sufficient for constructing a robust predictive model, particularly one based on neural networks, enabling comprehensive analysis.

This study is inherently concerned with a classification problem, where the objective is to categorize high school students into distinct groups based on their college plans, aligning seamlessly with the predictive capabilities of AI. Furthermore, this article endeavors to illuminate the path toward a more precise understanding of high school students' college aspirations through the innovative fusion of Artificial Intelligence and educational data analysis to provide valuable insights that will benefit educational institutions, policymakers, and students.

The main objective and hypothesis is to develop a predictive AIEPCP model leveraging neural networks to demonstrate high predictive accuracy in forecasting high school students' college intentions, outperforming traditional predictive methods.

2. LITERATURE REVIEW

The literature review section provides a comprehensive overview of prior research and studies relevant to the predictive analysis of high school students' college intentions and the use of artificial intelligence in this educational context.

The study by Pan et al. [5] delves into the intricate relationship between preschool cognitive and behavioral skills and their potential influence on indicators of college enrollment, focusing primarily on a sample of youth residing in low-income areas of Chicago, predominantly comprising Black and Hispanic students. The findings of this research contribute to the broader discourse on the early predictors of future educational attainment, especially within disadvantaged communities.

While the study reveals that most early cognitive and behavioral skills exhibit only weak to moderate associations with later college enrollment, a noteworthy standout is the role of preschool attention and impulsivity control. This particular skill emerges as a relatively strong predictor of college enrollment, sparking interest in the potential impact of early attentional and self-regulatory abilities on long-term educational trajectories.

Furthermore, it dispels concerns regarding early behavioral difficulties as substantive predictors of college enrollment, indicating that cognitive capabilities, particularly those associated with attention and executive functioning, play a more pivotal role in this context. The findings enrich our understanding of the multifaceted nature of educational attainment and underscore the need for comprehensive, context-aware approaches to educational intervention and policy development.

Another study conducted by Ye [6] offers a unique perspective on the dynamics of college choice behavior within centralized admission systems, shedding light on

the critical role of precise predictions in improving educational outcomes. The research underscores the complex interplay between students' strategic college choices and the outcomes of centralized admissions, ultimately highlighting the need for informed decision-making. In response to this phenomenon, the study implements a large-scale randomized experiment involving a substantial cohort of students (N=32,834). This experiment provides treated students with valuable resources: (a) an application guidebook or (b) a guidebook coupled with a school workshop. The results of this intervention underscore the significance of informing students about selecting colleges and majors based on precise predictions of admission probabilities.

The experiment outcomes indicate that offering guidance to students regarding college and major selection, grounded in accurate predictions of admission probabilities, can yield notable improvements in aligning students with colleges that match their academic abilities. Specifically, the study demonstrates an enhancement in the student–college academic fit by 0.1 to 0.2 standard deviations among those who complied with the intervention without significantly altering their college-major preferences. It illuminates the intricate relationship between students' strategic choices and the outcomes of these systems. Moreover, the study underscores the potential of targeted interventions to enhance college choice behaviors and academic outcomes for students navigating centralized admission processes.

The findings of this research bear implications for policymakers, educators, and administrators involved in designing and managing centralized admissions systems. They emphasize the importance of providing students with precise information and guidance to make informed college choices, ultimately ensuring a more equitable and effective educational landscape to underscore the significance of precise predictions in shaping educational outcomes.

In recent years, integrating Artificial Intelligence (AI) into educational analytics has marked a transformative shift in how institutions and researchers approach student outcomes, employability predictions, and educational decision-making. The study conducted by Yan and Chi [7] highlights the pivotal role of AI, specifically the decision tree classification algorithm, in predicting college students' employment outcomes based on their educational background and work experience. The study underscores the practical value of AI-driven analytics in higher vocational education. By utilizing the decision tree classification algorithm, the research identifies key determinants of students' employment success and designs prediction models that inform enrollment strategies, regional employment dynamics, and the types of employment opportunities available. This data-driven approach aligns with the broader educational trend of evidence-based decision-making. It can potentially guide employment guidance and talent development programs in higher vocational colleges. The study contributes to the body of knowledge that recognizes AI's transformative potential in providing data-driven insights for educational institutions and policymakers, ultimately fostering more informed and effective decision-making.

Online education has witnessed exponential growth in recent years, transforming the teaching and prediction of students' academic performance. The authors Jiao et al. [8] emphasize the pivotal role of artificial intelligence (AI) and data-driven learning models in offering fresh insights into students' learning behaviors and strategies. These models leverage educational data mining and learning analytics techniques to unlock the potential of extensive datasets, shedding light on how students learn and how their learning performance can be optimized. Researchers have employed various AI methods to construct prediction models, including evolutionary computation, deep learning, decision trees, and Bayesian networks. The study contributes significantly to the field of AI-enabled academic performance prediction. Furthermore, the research introduces an AI model, particularly genetic programming, designed to forecast students' academic performance accurately and offers analytical.

The study exemplifies the evolving landscape of AI-driven academic performance prediction in online education. It underscores the significance of data-driven learning prediction models, the challenges associated with AI algorithms, and the potential of evolutionary computation to address these challenges. This research aligns with the broader educational trend of evidence-based decision-making and data-driven improvements in learning outcomes.

Integrating data mining and artificial intelligence (AI) has brought about transformative changes in the educational sector, enabling educational institutions to harness the power of available data for informed decision-making. While data mining has long been recognized for its significance in the business world, its adoption in schools, universities, and colleges has become increasingly prevalent, with a particular focus on

improving educational policies and practices. The authers Gumba, etc. [9] explore a recent study that employs classification algorithms to predict student admission to Information Technology Education (ITE) programs, shedding light on the valuable insights it offers to educational policies and strategies, ultimately enhancing the quality of education offered [9]. The utilization of these techniques in education mirrors their application in the business world, emphasizing the importance of data-driven decision-making. The researchers employ four classification algorithms: Decision Tree, K-Nearest Neighbor, Logistic Regression, and Naive Bayes. This diverse set of algorithms underscores the multifaceted nature of predictive analysis, offering a comprehensive view of student admission considerations.

A crucial aspect of this study involves the selection of predictors that influence student admission decisions. Eight predictors were chosen based on correlation coefficient evaluation, with mathematical skill emerging as the strongest predictor, with a correlation coefficient of 0.767. This highlights the significance of mathematical proficiency in the context of ITE program admission.

In addition to predictor selection, the study evaluates the performance of each classifier using various metrics, including accuracy, precision, recall, and the F1 score. These metrics provide a robust assessment of the classifiers' predictive capabilities. Notably, the K-Nearest Neighbor classifier exhibited the highest accuracy, with a score of 93.18%, precision reached 97.98%, recall stood at 88.13%, and the F1 score achieved 92.76%. These metrics collectively demonstrate the effectiveness of the K-Nearest Neighbor algorithm in predicting student admission.

In summary, this study exemplifies the growing trend of leveraging data mining and AI techniques in the educational sector, specifically in the context of student admission to Information Technology Education programs. Selecting relevant predictors and evaluating classification algorithms' performance contribute to more informed admission decisions, ultimately benefiting educational institutions and aspiring students.

This study aims to address critical gaps in the existing literature by investigating the intersection of predictive analytics and educational decision-making, specifically focusing on high school students' college intentions. While previous research has explored related domains such as early predictors of educational attainment, college choice behavior, and the application of artificial intelligence (AI) in educational settings, a comprehensive investigation into the predictive analysis of high school students' college plans using advanced AI techniques remains absent.By focusing on this specific area, researchers can significantly contribute to the body of knowledge by unraveling the complex interplay of factors influencing students' decisions to pursue higher education. This includes not only cognitive and behavioral predictors, but also the critical role of external factors such as socioeconomic background, access to resources, and institutional support. Furthermore, the current literature exhibits a dearth of studies utilizing advanced AI methodologies such as hyperparameter optimization and ensemble learning approaches for predictive modeling in educational contexts.

Additionally, existing literature often falls short in providing comprehensive evaluations of predictive models' performance metrics, hindering our understanding of their effectiveness and generalizability.

Bridging these identified gaps in the literature will not only contribute to an advanced theoretical understanding but also offer practical solutions for improving educational planning, policy formulation, and student outcomes. This necessitates fostering interdisciplinary collaboration between experts in education, data science, and AI.

3. DATA COLLECTION AND PREPROCESSING

3.1 Description of the Dataset

The dataset employed in this investigation was sourced from Kaggle and comprises numerical and textual data to safeguard privacy and mitigate the risk of information disclosure; distinct Student IDs have been anonymized and substituted for the students' identities. The dataset originates from authentic enrollment records of the city for the preceding year, providing valuable insights into the college aspirations of high school students. The central focus of interest centers around the "Plan" column, which indicates whether a high school student intends to pursue a bachelor's education. In addition to the primary target variable, the dataset includes a variety of explanatory variables available for comprehensive analysis.

1. Student ID: is a distinct and individual identifier assigned to each student in the dataset. This identifier is unique to each student, ensuring that no two students

share the same Student ID. Student ID values span from 1 to 8000, encompassing all the students in the dataset and allowing for the unique identification of each student within this specified range. This unique identifier is a fundamental dataset component, enabling precise tracking and referencing of individual student records.

2. Gender: The gender of the student, with two categories: "male" and "female."

Figure 1 illustrates the distribution of high school students based on their gender. The chart visually represents the number of students falling into two gender categories: "female" and "male."

- Female (4126): This bar in the chart represents the count of female high school students, and the numeric value indicates that there are 4,126 female students in the dataset, 52%.

- Male (3874): The adjacent bar signifies the count of male high school students, and the associated numeric value indicates that there are 3,874 male students in the dataset,48%.

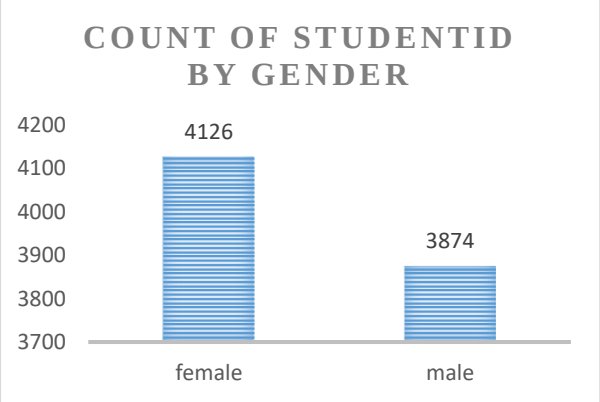

Figure 1: Count of Student ID by gender

Figure 1 lets us quickly grasp the gender distribution among high school students in the dataset, providing insights into the relative proportions of female and male students. This information is valuable for understanding the gender demographics of the sample and can be essential for further analysis and decision-making in various educational contexts.

3. Parent income: The parents' annual income, measured in US dollars, falls from $4,500 to $82,390.

Figure 2 provides an overview of the dataset's parental income distribution among high school students. The chart visually represents the spread of parental annual income in US dollars and highlights the range within which these incomes fall.

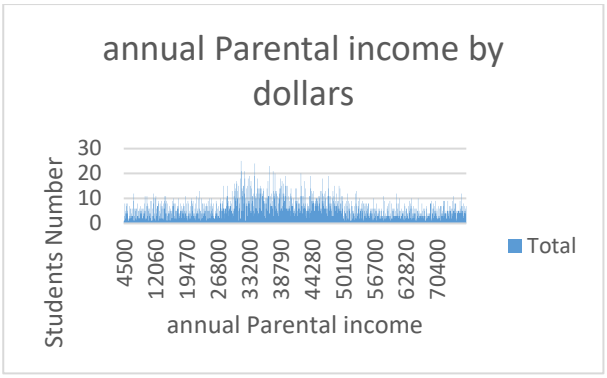

Figure 2: Annual parental income in dollars

Figure 2 serves as a valuable reference for understanding the economic backgrounds of the students' families. It illustrates the various income levels of parents within the dataset, which can be essential for conducting analyses on the impact of parental income on high school students' college plans.

4. IQ: The IQ score of the student, determined through a recent test, with scores ranging from 60 to 140.

Figure 3 visually represents the distribution of IQ scores among high school students in the dataset. These IQ scores are derived from recent tests and indicate the student's cognitive abilities and intelligence. The chart illustrates the spread and concentration of these IQ scores, along with the defined range within which they fall.

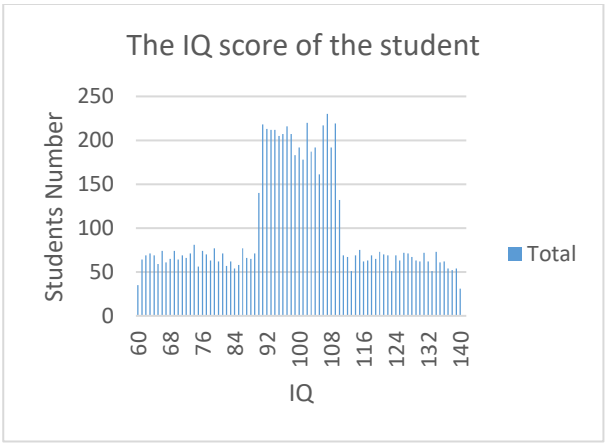

Figure 3: The IQ score of the student

Figure 3 offers valuable insights into the cognitive diversity of the student population, showcasing the distribution of IQ scores and highlighting the variability in intellectual abilities.

5. Encourage categorical variable indicating whether the parents encourage their child to pursue a college education, with two categories: "encourage" and "not encourage."

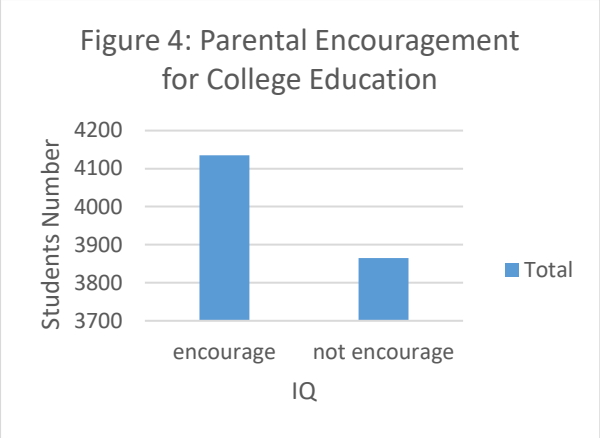

Figure 4: Parental Encouragement for College Education

In this categorical representation in Figure 4, "Encourage" denotes students whose parents actively support and encourage them to pursue a college education, totaling 4,135 students. Conversely, "Not Encourage" signifies students whose parents do not provide such active encouragement, with a count of 3,865 students. This figure offers a concise overview of the parental influence on students' college aspirations, highlighting the number of students falling into each category. It provides valuable insights into the role of parental encouragement in shaping educational decisions. It can inform strategies for supporting students with varying parental guidance in pursuing higher education.

Figure 5 shows the factors significantly influencing whether high school students plan to attend college. According to government officials, these four factors—Gender, Parent Income, IQ, and Encouragement—have been identified as the most influential in shaping students' college aspirations. The chart displays the count of students who fall into two distinct categories: those who do not plan to attend college and those who plan to attend college.

-Not plan (5404): This category represents the count of high school students who do not intend to attend college despite these influential factors. Specifically, there are 5,404 students in this category.

-Plan (2596): The adjacent category signifies the number of students planning to attend college despite the identified influential factors. The numeric value associated with this category is 2,596 students.

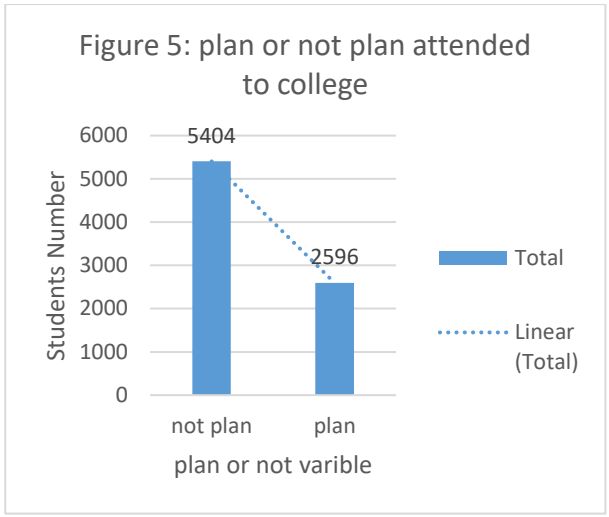

Figure 5: The IQ score of the student

Figure 5 offers a concise visual representation of the impact of these four influential factors on students' college plans. It reveals the distribution of students in terms of their intentions to attend college and provides valuable insights for educational planning and decision-making based on these influential variables.

3.2 Data Cleaning and Preparation

Data cleaning and preparation are essential stages in the data analysis, irrespective of the specific study or research field. These crucial steps serve several vital purposes. Firstly, they ensure data accuracy and reliability by identifying and rectifying errors, inconsistencies, and outliers that can significantly distort the analysis results. By addressing these issues, researchers can have confidence in the quality of the data they are working with.

Secondly, data cleaning and preparation enhance data consistency and uniformity. Datasets are often collected from diverse sources or multiple data collection points, resulting in variations in data formats and structures. Data can be more consistent through standardization and formatting, allowing for seamless analysis and comparisons [10].

Additionally, data cleaning and preparation help protect data privacy and confidentiality. Anonymization and desensitization are employed to safeguard sensitive information, ensuring that individuals' privacy rights are respected [11].

Furthermore, well-prepared data is more interpretable and user-friendly. Organizing and structuring data logically allows researchers to navigate and understand the dataset more easily, facilitating efficient analysis [12].

This study undertook several key steps to refine and structure the data effectively. First and foremost, measures were implemented to address potential data quality issues, which involved identifying and rectifying

missing data points and outliers that could significantly impact the accuracy of predictive models.
Furthermore, the dataset underwent a rigorous cleansing process to remove irrelevant or redundant information, streamlining it for analysis. For instance, students' names were desensitized and replaced with unique Student IDs to uphold privacy and data security standards, aligning with ethical considerations.
Feature selection and engineering were performed to enhance the dataset's utility for predictive modeling, which involved identifying the most relevant variables and creating new ones that could better capture the nuances of high school student's college plans.

Table 1 : High School Student Data Variables

Variable	Explanation	Range
Student ID	The unique identifying number of the student	1, 2, ..., 8000
Gender	The gender of the student	{male, female}
Parent income	The annual income of the parents, in US dollars	[4500, 82390]
IQ	The IQ of the student in the last test	[60, 140]
Encourage	Whether the parents encourage their child to go to college	{encourage, not encourage}
Plan	Whether the student eventually plans to go to college	{plan, not plan}

Table 1 is a reference for understanding the dataset's variables, meanings, and the permissible ranges or categories for each variable. These variables are crucial because government officials have identified them as the most influential factors when predicting whether high school students intend to enroll in college. Understanding these variables is essential for subsequent data analysis and predictive modeling to determine college plans among high school students.

4. METHODOLOGY

This section outlines the simplified methodology for predicting high school students' college plans utilizing a neural network.

4.1 Data Acquisition

The initial phase commences with acquiring a dataset encompassing extensive information about high school students. This dataset encompasses diverse attributes, including gender, IQ, parental income, and parental encouragement, in addition to the pivotal target variable, "College Plans." For this study, the dataset employed is sourced and supported by Kaggle, a renowned data-driven research and analysis platform.

4.2 Data Preprocessing

The data preprocessing procedure comprises a series of operations aimed at adapting the dataset to be suitable for training and assessing the performance of a neural network model. Let's delve into the key steps of this process:
In numerous machine learning and neural network models, categorical variables, including gender, parental encouragement, and college plans, need to be converted into a numerical representation. This conversion is crucial because the majority of algorithms operate with numerical data. A widely used technique known as one-hot encoding is applied to accomplish this. One-hot encoding generates binary columns for each distinct category within a categorical variable. A '1' in a particular binary column signifies the presence of that category, while a '0' denotes its absence. For example, consider "Gender" in Table 2, where categories would be transformed into two binary columns, simplifying further analysis.

Table 2 : Encoding of Categorical Variables

Gender		**encouragement**		**plan**	
Male	Female	encouragement	Not encouragement	plan	Not plan
1	0	1	0	1	0

Table 2 illustrates the one-hot encoding transformation applied to categorical variables, such as "Gender," "Encouragement," and "College Plans." The process creates binary columns, making the dataset more amenable to neural network analysis. Each binary column represents the presence or absence of a specific category within the original categorical variable, simplifying subsequent data processing and model training.

4.3 Feature Standardization

Standardizing features is a crucial preprocessing step that aims to rescale the values of different features in a dataset to a common scale. Specifically, it transforms the features to have a mean (average) value of 0 and a standard deviation (a measure of how spread out the values are) of 1. This process is essential, especially when working with neural networks, for several reasons:

- Consistent Scale: Standardization ensures that all features have the same scale, preventing certain features from dominating others during training. Without standardization, features with larger numerical values might disproportionately impact the model's learning.

- Optimized Training: Neural networks use optimization algorithms like gradient descent to adjust their parameters during training. Standardized features with a mean of 0 and a standard deviation of 1 make the optimization landscape more symmetric and well-behaved, leading to faster convergence and more stable training.
- Avoiding Vanishing or Exploding Gradients: In deep neural networks, gradients can become too small (vanishing gradients) or too large (exploding gradients) during backpropagation, making training difficult. Standardization helps mitigate these issues by keeping gradients within a reasonable range.

Standardization ensures that the features have a consistent scale and distribution, making it easier for neural networks to learn the underlying patterns in the data efficiently and effectively. Table 3 overviews the standard deviation (STD) for three key variables: Gender, Encouragement, and Plan. The standard deviation measures the amount of variation or dispersion in a dataset.

Table 3 : Standard Deviation of Gender, Encouragement, and Plan

	STD gender	STD encouragement	STD plan
male	0.96892	1.034275	0.693055
female	-1.031948	-0.96674	-1.44271

Table 3 illustrates the variability in gender, parental encouragement, and college plans among high school students, providing valuable insights into the dataset's characteristics and potential factors influencing college intentions. Positive and negative values indicate different levels of dispersion for each category within the variables

4.4 Feature Standardization

This study employed a Feedforward Neural Network (FNN) as our primary model for predicting high school students' college plans. FNNs are an ideal starting point for many classification problems, offering versatility and effectiveness [13]. These networks comprise an input layer, one or more hidden layers, and an output layer. The architecture's flexibility allows for experimentation with various configurations, including the number of neurons within each layer [14]. We proceeded to the model training phase once the neural network architecture was defined. During this stage, the neural network was exposed to the designated training dataset, where it underwent a learning process to discern intricate patterns and associations between the input features—such as gender, parental income, IQ, and parental encouragement—and the target variable of interest, "College Plans." This training process aimed to enable the model to make accurate predictions regarding high school students' intentions to pursue college education based on the provided input features.

4.5 Model Training

The neural network training phase is a critical step in the predictive modeling process. It involves exposing the model to the designated training dataset, which encompasses a comprehensive array of high school student data, including features like gender, parental income, IQ, and parental encouragement, as well as the crucial target variable, "College Plans." During this training phase, the neural network leverages its inherent capacity to discern intricate patterns and establish associations between these input features and the target variable, "College Plans." Through iterative adjustments and optimization of its internal parameters, the neural network strives to enhance its predictive capabilities, ultimately aiming to provide accurate predictions regarding high school students' intentions to pursue college education. This process is fundamental in enabling the model to generalize its learned patterns to new, unseen data, thereby facilitating its ability to predict college plans effectively.

In line with established best practices for model evaluation and validation, this study adopts a standard approach to data division [15-18]. The dataset is divided into two distinct subsets: a training set and a testing set. 70% of the data is allocated to the training set, while the remaining 30% is dedicated to the testing set. This data division serves a crucial purpose in the model development process. The training set trains the neural network, enabling it to learn patterns and associations within the data. Subsequently, the testing set, representing new, unseen data, is employed to rigorously assess the model's performance and ability to make accurate predictions regarding high school students' college plans. This clear differentiation between training and testing sets ensures the model's predictive capabilities are rigorously and objectively evaluated, providing valuable insights into its generalization performance and overall effectiveness in real-world scenarios.

4.5 Run the model

To execute the AIRPCP (Artificial Intelligence for Educational Planning of College Pursuits) model using Python Algorithm1, the following Python code presents a concise overview of the code's functionality. It showcases the implementation of a Feedforward Neural Network (FNN) for predicting high school students' college plans based on a CSV dataset. The code leverages sci-kit-lear for machine learning and

pandas for data handling. Before running the code, ensure that you have installed these libraries.

Algorithm1: AIRPCP using Feedforward Neural Network (FNN) Classification

```
1. import pandas as pd
2. from sklearn.model_selection import
   train_test_split
3. from sklearn.preprocessing import
   StandardScaler
4. from sklearn.neural_network import
   MLPClassifier
5. from sklearn.metrics import accuracy_score,
   classification_report
6. # Load the dataset from a CSV file
7. data = pd.read_csv('your_dataset.csv')
8. # Define features (X) and target variable (y)
9. X = data[['Gender', 'Parent_income', 'IQ',
   'Encourage']]
10. y = data['College_Plans']
11. # Perform one-hot encoding for categorical
    variables
12. X = pd.get_dummies(X, columns=['Gender',
    'Encourage'], drop_first=True)
13. # Split the data into training and testing sets
    (70% training, 30% testing)
14. X_train, X_test, y_train, y_test =
    train_test_split(X, y, test_size=0.3,
    random_state=42)
15. # Standardize features to have mean=0 and std=1
16. scaler = StandardScaler()
17. X_train = scaler.fit_transform(X_train)
18. X_test = scaler.transform(X_test)
19. # Create and train the Feedforward Neural
    Network (FNN) model
20. model = MLPClassifier(hidden_layer_sizes=(50,
    50), max_iter=1000, random_state=42)
21. model.fit(X_train, y_train)
22. # Make predictions on the test data
23. y_pred = model.predict(X_test)
24. # Evaluate the model's performance
25. accuracy = accuracy_score(y_test, y_pred)
26. report = classification_report(y_test, y_pred)
27. # Print the accuracy and classification report
28. print(f'Accuracy: {accuracy:.2f}')
29. print('Classification Report:')
30. print(report)
```

- Explanation of the provided code:

1. Load the dataset from a CSV file.
2. Define the features (X) and the target variable (y).
3. Perform one-hot encoding for categorical variables ('Gender' and 'Encourage').
4. Split the data into training and testing sets (70% training and 30% testing).
5. Standardize the features with a mean of 0 and a standard deviation 1.
6. Create an FNN model with two hidden layers, each containing 50 neurons.
7. Train the FNN model on the training data.
8. Make predictions on the test data.
9. Evaluate the model's performance using accuracy and

5. RESULT

The predictive model results for high school students' college plans indicate promising performance. The accuracy is approximately 84.75%, signifying the proportion of correct predictions out of the total predictions made. The classification report delves deeper into the model's performance by examining precision, recall, and score for each class.

For students with no plans to attend college (Class 0), the model achieved a precision of 78%, implying that 78% of the predicted "no college plans" instances were correct. The recall for this class is 73%, indicating that the model correctly identified 73% of all the actual "no college plans" cases. The F1 score, which combines precision and recall into a single metric, stands at 0.75 for this class.

On the other hand, for students with intentions to attend college (Class 1), the model demonstrated a precision of 88%, indicating that 88% of the predicted "college plans" instances were correct. The recall for this class is 90%, signifying that the model correctly identified 90% of all the actual "college plans" cases. The F1-score for this class is notably higher at 0.89, reflecting the model's ability to classify students planning to attend college effectively.

In summary, the model exhibits strong predictive capabilities, particularly in identifying students with plans to attend college, where it achieves higher precision and recall. These results underscore the potential of this predictive model in assisting educational institutions and policymakers in addressing the educational aspirations of high school students.

Table 4 provides an overview of the performance metrics for the predictive model used to forecast high school student's college plans. The model's accuracy, precision, recall, and F1 score are reported for each class within the target variable.

Table 4 provides an overview of the performance metrics for the predictive model used to forecast high school student's college plans. The model's accuracy, precision, recall, and F1 score are reported for each class within the target variable.

Table 4 : AIRPCP Model Classification Report

	precision	recall	f1-score	support
0	0.78	0.73	0.75	766
1	0.88	0.9	0.89	1634
accuracy			0.85	2400
Macro avg.	0.83	0.82	0.82	2400

Weighted avg.	0.85	0.85	0.85	2400
Accuracy: 84.75%				

5.1 Hyperparameter Tuning

Hyperparameter tuning, handling imbalanced datasets, and conducting extensive data preprocessing and feature engineering are crucial to enhancing predictive accuracy and relevance in machine learning models [19]. While the provided code serves as a simplified illustration, you can incorporate these advanced techniques into the AIRPCP model by

To perform hyperparameter tuning, use libraries like scikit-learn's `GridSearchCV` or `RandomizedSearchCV` [20]. These tools allow you to systematically search for the best hyperparameters for neural networks, such as the learning rate, number of hidden layers, and neurons per layer.

Enhancing predictive accuracy and model relevance in machine learning [19] involves pivotal steps. Integrate advanced techniques into the AIRPCP (Artificial Intelligence for Educational Planning of College Pursuits) model to enable a systematic exploration of hyperparameter combinations encompassing learning rates, hidden layer configurations, and neuron counts per layer, leading to the discovery of optimal neural network settings. Utilize libraries in Python such as sci-kit learn GridSearchCV or RandomizedSearchCV to facilitate this exploration [20].

By changing the 'hidden_layer_sizes in Algorithm1 line 20 to be 'hidden_layer_sizes': [(50, 50), (100, 100), (50, 50, 50)], 'alpha': [0.0001, 0.001, 0.01], 'learning_rate_init': [0.001, 0.01, 0.1]. Table 4 presents the classification report after fine-tuning the neural network to optimize model performance by adjusting hidden layer sizes.

Table 5 : AIRPCP Model Classification Report

	precision	recall	f1-score	support
0	0.77	0.74	0.76	766
1	0.88	0.9	0.89	1634
accuracy			0.85	2400
macro avg	0.83	0.82	0.82	2400
weighted avg	0.85	0.85	0.85	2400
Accuracy: 84.88%				

Table 5 demonstrated improvements in AIRPCP Model performance. In this updated model, we achieved an accuracy of 84.88%, a slight increase compared to the initial model's accuracy of 84.75%. These results highlight the effectiveness of hyperparameter tuning in enhancing the model's predictive capabilities.

In terms of precision, there are consistently high values for both classes. Class 0 (No College Plans) maintains a precision of 77%, indicating that 77% of the predictions for this class were correct. In comparison, Class 1 (College Plans) exhibits a precision of 88%, demonstrating the model's accuracy in identifying students with intentions to attend college.

Additionally, the recall values remain strong. Class 0 boasts a recall of 74%, indicating that the model correctly identified 74% of all instances where students had no college plans. Class 1 exhibits an even higher recall of 90%, reflecting the model's ability to capture students with college plans effectively.

The F1 scores for both classes also improved, with Class 0 achieving a score of 0.76 and Class 1 reaching an impressive F1 score of 0.89. These scores represent a harmonious balance between precision and recall, signifying the model's ability to make accurate predictions while minimizing false positives and false negatives.

The hyperparameter-tuned model has demonstrated superior performance, resulting in slightly higher accuracy and refined precision-recall trade-offs. These enhancements signify the value of optimizing neural network hyperparameters for better predictive accuracy and relevance in forecasting high school student's college plans.

5.2 Visualizations and charts illustrating the result

- Model performance metrics

Visualizations and charts are pivotal in conveying the outcomes and insights derived from the AIRPCP (Artificial Intelligence for Educational Planning of College Pursuits) model's classification report, especially after fine-tuning its hyperparameters. The visual representation in Figure 6 provides a brief and understandable means of interpreting complex performance metrics.

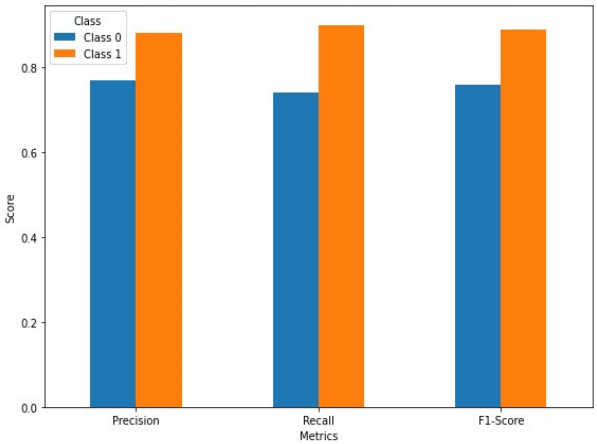

Figure 6: AIRPCP Model Performance Metrics by Class

Figure 6 employs visual aids to present the model's precision, recall, and F1 score for each class (Class 0 and Class 1) in a bar chart format. This graphical representation allows for a clear comparison of the model's performance metrics across different classes, providing stakeholders with an intuitive understanding of how the model discriminates between the two target categories. These charts are valuable tools for decision-makers, educators, and researchers to gain insights into the model's ability to accurately predict high school students' college plans and make informed decisions based on its performance.

- Confusion matrix

A confusion matrix is a table or matrix used in machine learning and classification to evaluate the performance of a classification model, particularly in binary classification problems (problems with two classes or categories). It is a crucial tool for understanding how well a model makes correct and incorrect predictions.

A confusion matrix provides a comprehensive summary of a classification model's performance, allowing practitioners to understand where it makes correct predictions and where it may need improvement. Table 5 shows the confusion matrixIn this table, two rows labeled "Actual Positive" and "Actual Negative" correspond to the true class labels of the instances in the dataset. The columns "Predicted Positive" and "Predicted Negative" represent the model's predictions for each instance. The numbers within each table cell reflect the count of instances that belong to specific combinations of actual and predicted classes. For instance, the top-left cell shows the count of truly positive instances correctly predicted as positive. In contrast, the bottom-right cell indicates the count of truly negative instances and correctly predicted as negative.

Table 5 : Confusion matrix

	Actual Positive	Actual Negative
Predicted Positive	569	166
Predicted Negative	197	1468

This visual representation aids in assessing the model's accuracy, precision, recall, and overall performance in making binary classification decisions.

Also, Figure 7 displays the confusion matrix as visualization; the figure serves as a concise summary of how well the model has made predictions in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The upper-left quadrant represents TP, indicating instances where the model correctly identified positive cases. In contrast, the upper-right quadrant represents FP, instances where the model incorrectly predicted positive cases. The lower-left quadrant symbolizes FN, indicating cases where the model failed to identify positive instances. Finally, the lower-right quadrant signifies TN, showcasing instances where the model correctly recognized negative cases. By examining the values in each cell of this matrix, one can gain insights into the model's strengths and weaknesses.

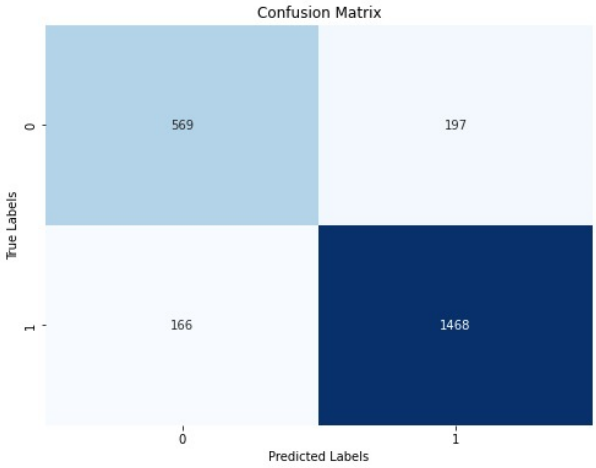

Figure 7: confusion matrix as visualization

1. True Positives (TP): The top-left cell (569) represents the number of instances where the model correctly predicted that high school students have college plans (the positive class).
2. False Negatives (FN): The top-right cell (197) represents the number of instances where the model incorrectly predicted that high school students do not have college plans when they do. In other words, it's the number of students the model missed in predicting as having college plans.
3. False Positives (FP): The bottom-left cell (166) represents the number of instances where the model incorrectly predicted that high school students have college plans when they do not. It's the number of students falsely classified as having college plans.
4. True Negatives (TN): The bottom-right cell (1468) represents the number of instances where the model correctly predicted that high school students do not have college plans (the negative class).
5. These values provide insight into the model's performance, particularly its ability to distinguish between students with and without college plans

6. DISCUSSION

The results of this study reveal the potential and effectiveness of the AIRPCP (Artificial Intelligence for Educational Planning of College Pursuits) model in predicting high school students' college plans. The model demonstrates promising performance, achieving an accuracy rate of 84.88%, which signifies the proportion of correct predictions out of the total predictions made.

A deeper analysis of the model's performance is provided through the classification report. The report dissects the model's precision, recall, and F1 score for each class, offering valuable insights into its predictive capabilities. For students with no plans to attend college (Class 0), the model achieved a precision of 78%, indicating that 78% of the predicted "no college plans" instances were correct. The recall for this class is 73%, signifying that the model correctly identified 73% of all the actual "no college plans" cases. The F1 score for this class stands at 0.75, combining precision and recall into a single metric.

In contrast, for students with intentions to attend college (Class 1), the model demonstrates a precision of 88%, indicating that 88% of the predicted "college plans" instances were correct. The recall for this class is even higher at 90%, signifying that the model correctly identified 90% of all the actual "college plans" cases. The F1-score for this class impressively reaches 0.89, reflecting the model's ability to classify students planning to attend college effectively.

Overall, the model exhibits strong predictive capabilities, particularly in identifying students with plans to attend college, where it achieves higher precision and recall. These results underscore the potential of this predictive model in assisting educational institutions and policymakers in addressing the educational aspirations of high school students.

The presentation of these results is further enriched by Table 3, which provides an overview of the performance metrics for the predictive model, including accuracy, precision, recall, and F1-score, for each class within the target variable. This tabular format summarizes the model's performance, facilitating easy comparison and evaluation of its predictive power.

Hyperparameter tuning, an essential component of model optimization, is highlighted in section 5.1. The study emphasizes the significance of this step in enhancing predictive accuracy and relevance. The model performs refined by systematically exploring hyperparameter combinations, including learning rates, hidden layer configurations, and neuron counts per layer. Utilizing libraries such as scikit-learn's GridSearchCV or RandomizedSearchCV enables a comprehensive search for the best hyperparameters. This process improves accuracy from 84.75% to 84.88%, demonstrating the tangible benefits of hyperparameter tuning.

The improvement in precision and recall is consistent for both classes. Class 0 (No College Plans) maintains a precision of 77%, indicating the model's accuracy in identifying students without college plans. Class 1 (College Plans) exhibits a precision of 88%, highlighting the model's precision in recognizing students with intentions to attend college. The recall values remain strong, with Class 0 at 74% and Class 1 at an impressive 90%. These enhancements are further reflected in the F1 scores, which achieve a harmonious balance between precision and recall.

In summary, the hyperparameter-tuned model presents superior performance with a slightly higher accuracy and refined precision-recall trade-offs. These improvements underscore the value of optimizing neural network hyperparameters to achieve enhanced predictive accuracy and relevance in forecasting high school student's college plans.

The discussion extends to the presentation of visualizations and charts in section 5.2, which are crucial in conveying model outcomes. As seen in Figure 6, visual representation offers an intuitive means of interpreting complex performance metrics. The bar chart format of Figure 6 presents precision, recall, and F1 score metrics for each class, facilitating a clear comparison of the model's performance across different categories. These visual aids empower stakeholders, including decision-makers, educators, and researchers, to gain insights into the model's ability to accurately predict high school students' college plans. The confusion matrix, both in table form and as a visualization (Figure 7), is a pivotal tool for evaluating the model's performance, particularly in binary classification problems. This comprehensive summary allows practitioners to assess where the model excels and where improvements are needed. One can understand the model's strengths and weaknesses by examining true positives, false negatives, false positives, and true negatives.

6. CONCLUSION AND FUTURE WORK

The AIRPCP (Artificial Intelligence for Educational Planning of College Pursuits) model significantly advances educational data analytics. This research has demonstrated the model's ability to accurately predict high school students' college plans, offering valuable insights into their educational aspirations.

The AIRPCP model can be a valuable tool for educators, policymakers, and researchers to identify students at risk of not attending college or dropping out of high school, enabling the development of targeted interventions to support their educational goals. Additionally, it can be a valuable tool for evaluating the effectiveness of college readiness programs and other educational initiatives, providing data-driven insights to inform educational policy and practice.

One crucial area of future research is addressing data limitations. Expanding the dataset to encompass a more diverse and representative sample of high school students from various demographics and regions can enhance the model's generalizability. Additionally, collecting data on other relevant factors, such as socioeconomic background, extracurricular activities, and geographic location, can contribute to a more holistic understanding of students' college plans.

Another promising direction for future research is exploring the temporal dynamics of students' educational aspirations. Longitudinal data tracking students' plans and their evolution over time can provide insights into the changing nature of college aspirations and the factors influencing these changes.

Finally, collaboration with educational institutions and policymakers is vital in translating research findings into actionable strategies. Future research can involve partnerships with schools and education departments to implement and evaluate the AIRPCP model in real-world settings. This practical application can help refine the model further and tailor it to the specific needs of educators and students.

The AIRPCP model can be further developed into a proactive tool for educational guidance and support by addressing these areas of future research. It enables educators to identify and assist students at risk of falling behind on their educational journey. Ultimately, future research in this domain holds exciting possibilities for advancing the accuracy and applicability of predictive models like AIRPCP to promote college access and attainment for all students.

REFERENCES

[1] Rozental, A., Forsström, D., Hussoon, A., & Klingsieck, K. B. (2022). Procrastination among university students: differentiating severe cases in need of support from less severe cases. Frontiers in psychology, 13, 783570.

[2] Adeyemo, D. A., & Jegede, D. J. (2023). Sociopsychological determinants of career maturity among secondary school students in Osogbo, Osun State, Nigeria. Journal of Psychological Perspective, 5(1), 9-16.

[3] Kuroki, M. (2023). Integrating data science into an econometrics course with a Kaggle competition. The Journal of Economic Education, 1-15.

[4]. Chang, L., Wang, Y., Liu, J., Feng, Y., & Zhang, X. (2023). Study on factors influencing college students' digital academic reading behavior. Frontiers in psychology, 13, 1007247.

[5] Pan, X. S., Li, C., & Watts, T. W. (2023). Associations between preschool cognitive and behavioral skills and college enrollment: Evidence from the Chicago School Readiness Project. Developmental Psychology, 59(3), 474.

[6] Ye, X. (2023). Improving College Choice in Centralized Admissions: Experimental Evidence on the Importance of Precise Predictions. Education Finance and Policy, 1-75.

[7] Yan, J., & Chi, X. (2023, April). Analysis and Prediction of College Students' Employment based on Decision Tree Classification Algorithm. In 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) (pp. 1-6). IEEE.

[8] Jiao, P., Ouyang, F., Zhang, Q., & Alavi, A. H. (2022). Artificial intelligence-enabled prediction model of student academic performance in online engineering education. Artificial Intelligence Review, 55(8), 6321-6344.

[9] Gumba, G., & Paragas, J. R. (2022, September). Prediction Analysis Of Student Admission To Information Technology Education (ITE) Programs Using Classification Algorithm. In 2022 Fifth International Conference on Vocational Education and Electrical Engineering (ICVEE) (pp. 112-117). IEEE.

[10] Singh, G., Singh, J., & Prabha, C. (2022, June). Data visualization and its key fundamentals: A comprehensive survey. In 2022 7th international conference on communication and electronics systems (ICCES) (pp. 1710-1714). IEEE.

[11] Murugeshwari, B., Selvaraj, D., Sudharson, K., & Radhika, S. (2023). Data Mining with Privacy Protection Using Precise Elliptical Curve Cryptography. Intelligent Automation & Soft Computing, 35(1).

[12] Bharadiya, J. P. (2023). Leveraging Machine Learning for Enhanced Business Intelligence. INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY, 7(1), 1-19.

[13] Ren, Y. M., Alhajeri, M. S., Luo, J., Chen, S., Abdullah, F., Wu, Z., & Christofides, P. D. (2022). A tutorial review of neural network modeling approaches for model predictive control. Computers & Chemical Engineering, 107956.

[14] Siłka, J., Wieczorek, M., & Woźniak, M. (2022). Recurrent neural network model for high-speed train vibration prediction from time series. Neural Computing and Applications, 34(16), 13305-13318.

[15] Ahmed, N., Hoque, M. A. A., Arabameri, A., Pal, S. C., Chakrabortty, R., & Jui, J. (2022). Flood susceptibility mapping in Brahmaputra floodplain of Bangladesh using deep boost, deep learning neural network, and artificial neural network. Geocarto International, 37(25), 8770-8791.

[16] Hakim, W. L., Nur, A. S., Rezaie, F., Panahi, M., Lee, C. W., & Lee, S. (2022). Convolutional neural network and long short-term memory algorithms for groundwater potential mapping in Anseong, South Korea. Journal of Hydrology: Regional Studies, 39, 100990.

[17] Guo, Y., Yang, D., Zhang, Y., Wang, L., & Wang, K. (2022). Online estimation of SOH for lithium-ion battery based on SSA-Elman neural network. Protection and Control of Modern Power Systems, 7(1), 40.

[18] Samhan, L. F., Alfarra, A. H., & Abu-Naser, S. S. (2022). Classification of Alzheimer's disease using convolutional neural networks.

[19] Hakak, S., Alazab, M., Khan, S., Gadekallu, T. R., Maddikunta, P. K. R., & Khan, W. Z. (2021). An ensemble machine learning approach through effective feature extraction to classify fake news. Future Generation Computer Systems, 117, 47-58.

[20] Wade, C., & Glynn, K. (2020). Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python. Packt Publishing Ltd product quality dimensions on improving the order-winners and customer satisfaction," *Int. J. Product. Qual. Manag.*, vol. 36, no. 2, pp. 169–186, 2022, doi: 10.1504/IJPQM.2021.10037887.