UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Career-Driven Study Pathways Prediction for Secondary School Students Using Hybridized Machine Learning

Muhammad Hanafi Thman, Hasyiya Karimah Adli,
Mohd Hakimi Aiman, Hadhrami Ab. Ghani* & Normasliza Malik
Faculty of Data Science and Computing
Universiti Malaysia Kelantan
Pengkalan Chepa, Malaysia
Email: s24e0359f@siswa.umk.edu.my; hasyiya@umk.edu.my; aiman.i@umk.edu.my; hadhrami.ag@umk.edu.my; s24e0359f@siswa.umk.edu.my

Submitted: 15/9/2025. Revised edition: 28/11/2025. Accepted: 5/1/2026. Published online: 28/7/2026
DOI: https://doi.org/10.11113/ijic.v16n1-2.698

***Abstract*—Future study pathway determination is paramount for the secondary students' further academic development especially in preparing themselves for their careers. Without sufficient information and guidance, students might be exposed to information-limited advice and choices which tend to risk their future undertaking. Addressing this crucial decision-making challenge in choosing the study pathway while considering the students' career dreams, this paper proposes hybridized machine learning models which are career-driven in guiding the students to make the right study pathway decision. Upon pre-processing the dataset containing essential students' academic performance-related data as well as their career aspirations, four different hybridized machine learning models are developed to be trained and tested with the data for predicting the students' study pathways. The machine learning models studied include random forests, decision trees as well as neural networks which are hybridized between each other to benefit the diversifying strengths of each model. The performance metrics such as accuracy and precision have been measured cross-validated to verify the model's dependability. The obtained results demonstrated that the proposed career-driven hybridized models perform better than the typical existing machine learning models, which echo into better decisions offered by the proposed hybridized models in advising students' future study pathway more effectively.**

***Keywords*—Study pathway recommendation, Machine learning, Data-driven education, Predictive modeling, Student achievement**

I. INTRODUCTION

Choosing the right study pathway remains a critical challenge for many secondary school students, particularly those in their final years of schooling at the secondary school level. At this stage, students must select appropriate pathways such as matriculation, foundation, or diploma programs that will significantly influence their academic progression and career prospects [1].

This decision-making process is complex, as it depends on multiple factors including academic performance, personal interests, family background, and long-term career aspirations. Poorly informed choices often result in reduced motivation, academic underperformance, and dissatisfaction in future career trajectories.

Consequently, many students select educational pathways that are poorly aligned with their abilities, personal interests, and long-term career objectives [2]. Such mismatches often lead to diminished motivation, lower academic achievement, and dissatisfaction with future career outcomes. In addition, the increasing diversity of student profiles and the growing complexity of academic and occupational environments highlight the need for more sophisticated and personalized approaches to study pathway guidance.

Traditionally, schools provide career counseling to assist students in making these important decisions. Career counseling is intended to help learners understand their strengths and align them with suitable academic and career opportunities. However, conventional counseling approaches are often generalized and resource-constrained, making it difficult to provide personalized guidance for each student. This lack of tailored support contributes to suboptimal decisions that may affect long-term educational outcomes.

The emergence of data science with machine learning methods provide opportunities to address this gap through personalized and data-driven recommendations. By analyzing patterns in academic records, interests, and career aspirations, machine learning can generate tailored study pathway suggestions that align with each student's abilities and goals [3]. Such approaches have already been applied in higher education to predict dropout risks, improve retention, and recommend personalized learning strategies [4]. In addition, recommender approaches have been widely adopted in other domains such as e-commerce and entertainment, demonstrating their potential to deliver customized experiences [5]. However, predicting study pathways driven by the students' career aspirations aided by suitably hybridized machine learning remain as a significant problem that has not been addressed in literature

Within education, study pathway recommendation can provide students with individualized guidance on course and program selection, thereby increasing engagement, motivation, and academic success. They can also serve as early warning signals to identify at-risk students, enabling timely interventions. Moreover, educational institutions can leverage such approaches to optimize resource allocation in academic advising and support services. Despite these benefits, limited work has been conducted on applying data-driven recommendation approaches specifically for secondary school students' study pathway selection, particularly in local contexts.

This study aims to bridge this gap by developing a data-driven recommendation approach tailored to secondary school students. The proposed recommendation approach leverages machine learning techniques to analyze academic performance, interests, and career aspirations in order to suggest the most suitable study pathways. The main contributions of this work are as follows:

1) Design and development of a career-driven study pathway recommendation approach that integrates data science and hybridizes suitable machine learning techniques to generate personalized study pathway suggestions for secondary school students.
2) Analysis of key influencing factors—including academic performance, personal interests, and career goals—that shape students' pathway decisions.
3) Evaluation of the proposed approach's effectiveness using standard machine learning metrics (accuracy, precision, recall, and F1-score), alongside feedback from educators and students, to validate its relevance and usability.

By addressing the limitations of traditional guidance approaches, this research contributes to the field of educational technology and demonstrates the potential of data-driven approaches in supporting students' academic and career decision-making. In the next section, the literature review is presented, followed by the research method and the result sections, before concluded by the conclusion section.

II. LITERATURE REVIEW

The integration of data science into education has evolved from descriptive analytics to advanced predictive modeling that supports academic planning and improvement. Recent studies demonstrate that machine learning can significantly enhance the accuracy of performance forecasting and guidance in STEM and higher education contexts [6,7]. Such models enable institutions to anticipate learning outcomes and identify students requiring additional support, shifting from reactive to proactive educational decision-making [8,9].

Research on educational recommender systems further reveals how predictive analytics can inform individualized study pathways. [1] developed machine learning models to predict students' study path selection, showing that behavioral and contextual data contribute to improved recommendation accuracy. Similarly, [10] emphasized the role of data science in career path analysis, highlighting its value in aligning learning choices with employability trends. These approaches demonstrate that data-driven insights can connect academic trajectories with long-term career aspirations.

The selection of key predictive variables remains a critical factor in model design. Studies commonly incorporate academic performance indicators, socioeconomic status, and personal interests as major determinants of educational outcomes [11,12]. Socioeconomic variables, in particular, influence access to learning resources and academic motivation [13]. It has been also found in [14] that family expectations and perceived social mobility affect students' higher education choices, reinforcing the need to integrate contextual and psychological data into predictive frameworks.

Despite these advances, challenges persist in model generalization and ethical deployment. It has been observed in [4,15] that data imbalance and limited interpretability hinder the reliability of prediction systems. Concerns about transparency, bias, and data privacy remain central, especially when models are applied to sensitive student information. Addressing these issues requires balanced datasets, explainable algorithms, and adherence to responsible data governance practices.

Collectively, prior research underscores the growing maturity of data-driven educational analytics but also highlights gaps in ethical assurance, contextual adaptability, and variable interpretability. Building upon these findings, the present study seeks to develop an integrative framework that enhances prediction accuracy while ensuring fairness and accountability in educational guidance systems.

A. Recommendation Approaches in Education

In the educational domain, recommendation approaches have emerged as a critical tool for providing personalized guidance and supporting informed academic decision-making. By leveraging artificial intelligence, advanced analytics, and machine learning techniques, these approaches are capable of processing vast amounts of educational data to generate individualized suggestions for students. Such recommendations can extend from identifying suitable courses and learning resources to offering guidance on long-term academic and career trajectories. Recent developments in this field highlight their potential to enhance student engagement, improve academic outcomes, and facilitate more informed educational planning.

One of the primary applications of these approaches lies in the personalization of learning resources. Algorithms can analyze students' prior performance, preferred learning styles, and interests to recommend relevant courses, reading materials, online platforms, and even extracurricular activities. This approach not only addresses knowledge gaps but also promotes deeper engagement by aligning learning experiences with students' strengths and aspirations. As a result, learners benefit from more efficient and meaningful educational pathways.

A second area of application relates to pathway guidance, particularly for students at pivotal transition stages, such as those preparing to advance from secondary school to higher education. By integrating variables such as academic achievement, prior course history, individual preferences, and career ambitions, study recommendations can suggest routes that align with learners' long-term goals [11]. This function is especially relevant for senior secondary students who face complex decisions about their future.

Study recommendations are also increasingly being integrated into early warning frameworks. Through the detection of patterns in attendance, grades, and behavioral indicators, these approaches can identify students at risk of underperformance or dropout. Educators can then intervene proactively with tailored support, thereby enhancing retention and ensuring equitable opportunities for success.

Despite these benefits, the deployment of study recommendations in education raises several ethical and technical challenges. Concerns related to privacy, transparency, and algorithmic bias must be addressed to ensure that such tools are used responsibly. Protecting sensitive student data and providing clarity on how recommendations are generated are essential to fostering trust and equity in their use.

Overall, study pathway recommendations represent a transformative development in education, enabling personalized support and evidence-based decision-making. However, their effectiveness depends on the responsible integration of data science techniques and careful attention to ethical considerations.

B. Review on Factors Influencing Pathway Choices

The effectiveness of any recommendation approach in education is inherently tied to its ability to capture and interpret the diverse factors shaping students' academic choices. Research consistently emphasizes that study pathway selection is a multifaceted process, influenced by academic performance, personal interests, socioeconomic conditions, parental and peer influence, and the broader educational environment.

Students' study pathway decisions are shaped by a complex interplay of academic, personal, social, and institutional factors. Academic performance remains a major determinant, as students with stronger grades often pursue more competitive and demanding programs, while those with lower achievement levels are guided toward vocational or alternative tracks [16]. Alongside achievement, personal interests play a crucial role students are more motivated and perform better when their studies align with their passions and long-term career aspirations [17]. These individual choices, however, are also influenced by socioeconomic realities, as students from higher-income families tend to enjoy wider opportunities and support systems compared to those from less privileged backgrounds, emphasizing the need for equitable recommendation systems [12,13].

Beyond individual and economic aspects, parental and peer influences significantly affect students' academic directions, with family expectations and social networks shaping perceptions of available options [14,15,18]. The educational environment further strengthens or limits these choices institutions with quality teachers, robust guidance services, and career counseling help students make more informed decisions. Recent advances in artificial intelligence (AI) and data science have enhanced this process by enabling predictive analytics for early intervention and personalized learning support [19]. However, while AI-driven tools improve adaptability and institutional efficiency, they also introduce challenges of fairness, privacy, and ethical governance that must be carefully managed in future implementations.

While prior studies elucidate the multifaceted determinants of academic pathway selection, few integrate these psychosocial, economic, and institutional factors within a unified AI-driven framework. Current research often isolates predictors academic, social, or technological without exploring their combined predictive synergy. Thus, future studies should emphasize hybrid, ethically grounded AI models capable of balancing personalization, fairness, and explainability in digital education contexts.

C. Academic Performance Prediction

The prediction of academic success has been a central theme in recent research, with machine learning approaches showing strong potential. Ensemble methods, genetic algorithms, and feature engineering have been applied to identify at-risk students and optimize instructional strategies [20-22].

Evidence suggests that models such as Random Forest and boosted trees perform particularly well in predicting student outcomes across various contexts [6,11]. These studies collectively demonstrate the effectiveness of predictive analytics in supporting student success, while also drawing attention to the ethical challenges of deploying such methods in educational settings.

D. Machine Learning for Educational Prediction

Machine learning (ML) has become integral to transforming raw educational data into actionable insights. Recent research shows that ML-driven recommendation methods can analyze academic records, learning behaviors, and personal preferences to construct personalized study pathways [15]. Such approaches optimize learning by identifying strengths and weaknesses, recommending targeted interventions, and aligning study plans with students' long-term goals.

As tabulated in Table I, the comparative results presented highlight notable variations in accuracy across different studies. A performance of 98.15% accuracy has been reported in [23], which demonstrates the robustness of their proposed model in handling classification tasks. An accuracy of 94.7% has been reported in [17], reflecting strong yet comparatively lower effectiveness under similar evaluation conditions. In contrast, an accuracy of 89.6% is reported in [24], suggesting that earlier approaches were less optimized compared to recent advances.

Interestingly, [9] attained 96.3% accuracy, positioning their work between Wang and [17] and [23], thereby indicating progressive methodological improvements in recent years.

TABLE I. MACHINE LEARNING IN STUDENT PERFORMANCE PREDICTION

Related Work	Objective	Algorithm	Results/Findings
[7]	School performance ML	MLP classifier	86–99% accuracy
[17]	Online learning model	Logistic Regression	94.7%
[9]	Hybrid model study	CNN+Attention+XGB	96.3%
[23]	Student performance prediction	Random Forest	98.15%
[25]	Algorithm comparison	RF, SVM, NN	NN best
[26]	Multi-class grading	KNN, SVM, RF	2-class best
[8]	Methods survey	ANN, SVM, KNN	ANN best
[24]	Programming course ML	M5P, Linear Reg.	LRC best
[27]	ML usage overview	Regression, SVM	60–80% accuracy
[28]	Feature-based prediction	C4.5, SMO, 1-NN	Decision tree

Systematic reviews indicate that academic performance prediction often relies on diverse input variables, including personal, demographic, and behavioral data [8]. Algorithms such as random forests, support vector machines, and neural networks have consistently demonstrated high predictive accuracy [25]. Other approaches, including feature selection and ensemble methods, have also been applied to enhance efficiency and reliability [10,28].

Collectively, these studies underscore the transformative potential of machine learning in education, while also stressing the importance of addressing fairness, interpretability, and data governance in the design of such approaches.

III. RESEARCH METHOD

This study adopts a systematic methodology to design and evaluate a data-driven recommendation approach aimed at guiding secondary school students in selecting suitable academic pathways. The approach integrates survey-based data collection with secondary datasets, data preprocessing, exploratory analysis, machine learning model development, and validation. Each stage was designed to ensure reliability, reproducibility, and ethical handling of student information.

A. Methodological Framework

The proposed career-drive study pathway prediction framework follows a concise data science life cycle, as shown in Fig. 1, which involves four core phases: problem definition, data preparation, modeling, and evaluation. The problem definition stage focused on identifying the academic and career decision-making challenges faced by secondary school students. Data were then collected from two main sources: (i) a structured Google Form survey targeting students across different schools, and (ii) the publicly available Student Performance Dataset from Kaggle.

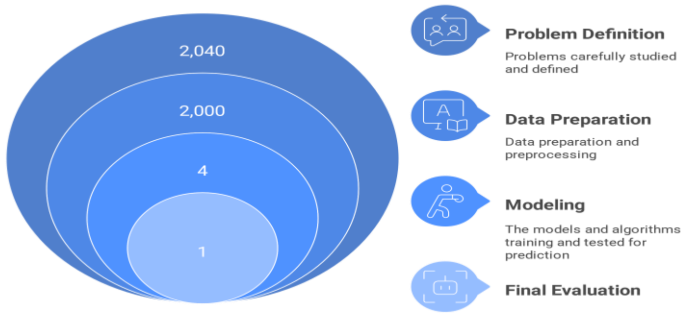

Fig. 1. Career-Driven Study Pathway Prediction Framework

As visualized in Fig. 1, 2,040 records were identified during problem definition, which were then cleaned and reduced to 2,000 valid samples in the data preparation stage. From these, four hybrid models were developed and tested during the modeling phase, leading to the selection of one final, optimized model for evaluation and validation.

B. Data Preparation

The survey instrument was designed with six sections to capture diverse factors influencing pathway selection: demographics, academic achievement, parental involvement, socioeconomic background, career aspirations, and external influences such as peer or teacher guidance. Forty students participated, providing primary contextual data. The Kaggle dataset, which is summarized in Table II, contributed a broader set of 2,000 student records containing demographic attributes, subject scores, extracurricular participation, and self-reported aspirations. Together, these datasets provided both localized insights and generalizable patterns.

TABLE II. STUDENTS' PERFORMANCE DATASET

Column	Description
id	A unique identifier for the student.
first_name and last_name	Student name.
email	Student email address.
gender	Gender of the student.
part_time_job	Whether the student has a part-time job.
absence_days	Number of days the student was absent from school.
extracurricular_activities	If the student did extracurricular activities
weekly_self_study_hours	Time spent on self-study during a week.
career_aspiration	The student's career aspiration.
Subject Scores	Scores include mathematics, history, physics, chemistry, biology, and English, geography

Prior to analysis, datasets underwent cleaning, normalization, and encoding. Incomplete or duplicate entries were removed, categorical variables (e.g., gender, career choice) were converted into numerical codes, and continuous variables were normalized to ensure comparability across sources. Outliers were inspected and addressed to maintain data integrity. This process ensured that the datasets were consistent, reliable, and ready for computational analysis.

In this research paper, the gender column will be encoded as 0 for Male and 1 for female. Each numerical number represents at least one career aspiration. Based on the career aspirations chosen by the students, the dataset is imbalanced as most students aspire to be a software engineer (315), a business owner (309), a banker (169) and an accountant (126).

C. *Analytical Techniques*

The study applied both statistical analysis and machine learning methods. Descriptive statistics and exploratory data analysis (EDA) were employed to identify correlations and influential features. For predictive modeling, four algorithms were tested: Decision Trees, Random Forests, Support Vector Machines (SVM), and Neural Networks. Each was trained on a subset of the data and tested on a hold-out sample. Performance was assessed using accuracy, precision, recall, and F1-score, allowing for a balanced evaluation of predictive capability.

D. *Modelling*

The proposed recommendation model was developed by selecting algorithms with the best performance on evaluation metrics. This model was then implemented to generate personalized study-path recommendations tailored to student profiles. The approach was validated through repeated experiments to ensure robustness and generalizability.

Four hybrid models were designed and evaluated to enhance the prediction of students' recommended study pathways: a baseline Artificial Neural Network (ANN), Random Forest combined with ANN (RF+ANN), Support Vector Machine combined with ANN (SVM+ANN), and Gradient Boosting Machine combined with ANN (GBM+ANN). The ANN served as the baseline classifier, while the hybrid models integrated ensemble or margin-based learners to refine feature representations before final classification. Experimental results revealed that the SVM+ANN model achieved the highest accuracy, as the nonlinear separability and margin maximization of SVM improved the generalization capability of the ANN. Compared to RF+ANN and GBM+ANN, this hybridization provided more discriminative input features, thereby reducing misclassification and demonstrating stronger empirical validation. The novelty of this study lies in the tailored integration of classical machine learning models with ANN for educational pathway prediction, a domain where hybrid learning strategies remain underexplored, thus offering new insights into intelligent academic guidance approaches.

A core principle underlying the baseline Artificial Neural Network (ANN) is the forward propagation mechanism, expressed as

$$a^{(l)} = f\left(W^{(l)}a^{(l-1)} + b^{(l)}\right), a^{(l)} = f\left(W^{(l)}a^{(l-1)} + b^{(l)}\right), \quad (1)$$

where $W^{(l)}W^{(l)}$ and $b^{(l)}b^{(l)}$ denote the weight matrix and bias vector at layer l, $a^{(l-1)}a^{(l-1)}$ represents the activations from the preceding layer, and $f(\cdot)f(\cdot)$ is a nonlinear activation function. This formulation enables the ANN to capture complex nonlinear mappings from input features to predicted outcomes, and it serves as the final classification stage in all four hybrid models.

Within the Random Forest (RF) framework, which underpins the RF+ANN hybrid model, feature representation is improved through ensemble averaging. The RF prediction rule can be written as

$$\hat{y} = \mathrm{majority_vote}\{h_t(x),\ t = 1,2,\dots,T\},$$
$$\hat{y} = \mathrm{majority_vote}\{h_t(x),\ t = 1,2,\dots,T\}, \quad (2)$$

here $h_t(x)h_t(x)$ denotes the prediction of the t t-th decision tree and TT is the total number of trees in the forest. This ensemble mechanism reduces variance and enhances feature discriminability, thereby providing more stable and informative representations for subsequent ANN-based classification. The integration of Support Vector Machines (SVM) within the SVM+ANN hybrid is guided by the margin maximization principle, formalized as

$$\frac{1}{2}|w|^2 \quad \text{subject to} \quad y_i(w \cdot x_i + b) \geq 1$$
$$\frac{1}{2}|w|^2 \quad \text{subject to} \quad y_i(w \cdot x_i + b) \geq 1, \quad (3)$$

where ww defines the separating hyperplane, bb is the bias term, and y_iy_i represents the class label of training instance x_i x_i. By enforcing large margins between classes, SVM produces refined, linearly separable feature representations that improve the generalization capability of the ANN. Empirical findings demonstrate that this margin-based refinement is central to the superior performance of the SVM+ANN hybrid. In contrast, the Gradient Boosting Machine (GBM) incorporated in the GBM+ANN hybrid relies on an additive learning process, expressed as

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x),$$
$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x), \quad (4)$$

where $F_m(x)F_m(x)$ denotes the boosted model at iteration mm, $h_m(x)h_m(x)$ represents a weak learner fitted to the residual errors of the previous stage, and $\gamma_m\gamma_m$ is the corresponding learning rate. This iterative refinement strengthens sensitivity to underrepresented classes and patterns in the data, thereby addressing imbalance issues. However, the sequential nature of boosting also increases susceptibility to overfitting, as evidenced in the comparative analysis against SVM+ANN.

IV. FINAL EVALUATION AND MEASUREMENT

The different metrics that can be used to measure the effectiveness of the recommendation approach include accuracy, precision, recall, and F1-score. Accuracy is defined as the ratio between correct recommendations and the total number of recommendations. Recall is basically the fraction of relevant recommendations out of all the potentially relevant recommendations, while precision is the fraction of relevant recommendations among the retrieved recommendations. F1-score is the harmonic means of precision and recall; it is a fair measure that describes the effectiveness of the proposed approach. The measurement metrics or parameters used in this research paper are as follows:

A. Accuracy

The accuracy of the multi-class classification for predicting the study pathway based on the careers is given as follows:

$$\text{Accuracy} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i+FP_i+FN_i)}$$

$$\text{Accuracy} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i+FP_i+FN_i)}, \quad (5) \quad (5)$$

where CC = number of classes, TP_iTP_i = true positives for class ii, FP_iFP_i = false positives for class ii, FN_iFN_i = false negatives for class ii.

B. Precision (per class i)

The precision of the multi-class classification for predicting the study pathway based on the careers is given as follows:

$$\text{Precision}_i = \frac{TP_i}{TP_i+FP_i} \text{Precision}_i = \frac{TP_i}{TP_i+FP_i},$$

(6) (6)

where the Macro-Precision and Micro-Precision are given as follows

Macro-Precision: $\frac{1}{C}\sum_{i=1}^{C} \text{Precision}_i \frac{1}{C}\sum_{i=1}^{C} \text{Precision}_i$

(7) (7)

Micro-Precision: $\frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i+FP_i)} \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i+FP_i)}$

(8) (8)

C. Recall (per class i)

The Recall formula of the multi-class classification for predicting the study pathway based on the careers is given as follows:

$$\text{Recall}_i = \frac{TP_i}{TP_i+FN_i} \text{Recall}_i = \frac{TP_i}{TP_i+FN_i},$$

(9) (9)

with

Macro-Recall: $\frac{1}{C}\sum_{i=1}^{C} \text{Recall}_i \frac{1}{C}\sum_{i=1}^{C} \text{Recall}_i$

(10) (10)

Micro-Recall: $\frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i+FN_i)} \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i+FN_i)}$

(11) (8)

D. F1-Scorel (per class i)

The F1-Score formula of the multi-class classification for predicting the study pathway based on the careers is given as follows:

$$F1_i = \frac{2\times\text{Precision}_i\times\text{Recall}_i}{\text{Precision}_i+\text{Recall}_i} F1_i = \frac{2\times\text{Precision}_i\times\text{Recall}_i}{\text{Precision}_i+\text{Recall}_i},$$

(12)

with

Macro-F1: $\frac{1}{C}\sum_{i=1}^{C} F1_i \frac{1}{C}\sum_{i=1}^{C} F1_i$

(13)

Micro-F1: $F1_{\text{micro}} = \frac{2\cdot\text{Precision}_{\text{micro}}\cdot\text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}}+\text{Recall}_{\text{micro}}}$

$F1_{\text{micro}} = \frac{2\cdot\text{Precision}_{\text{micro}}\cdot\text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}}+\text{Recall}_{\text{micro}}}$ (14)

V. Ethical Considerations

Ethical and regulatory considerations were integral to the development and implementation of the proposed AI-based educational and decision-support framework. All datasets were processed under strict adherence to established protocols of confidentiality and data protection, ensuring that no individual identity could be traced. Emphasis was placed on model transparency and interpretability to strengthen accountability and nurture users' confidence, especially when system-generated insights could affect academic progression or career pathways. The conceptual foundation of this framework resonates with international standards such as the General Data Protection Regulation (GDPR) and comparable legislative directives that advocate equitable data usage, responsible governance, and algorithmic clarity. Subsequent applications will incorporate systematic compliance evaluations and explicit consent procedures to reinforce the ethical integrity of AI adoption within authentic educational environments.

VI. Result

This study utilizes the *student_score* dataset as a foundation for constructing a recommendation framework guided by the data science life cycle. The process involves systematic data preparation, application of machine learning models, and rigorous performance evaluation The primary objective is to assess the predictive ability of various algorithms in identifying suitable academic pathways for secondary school students. Four distinct classifiers—Random Forest, Support Vector Machine (SVM), Decision Tree, and Neural Network—are implemented and benchmarked against one another. Their performance is measured using standard evaluation metrics, including accuracy, precision, recall, and F1-score, thereby enabling the identification of the most effective model to support students in making informed choices

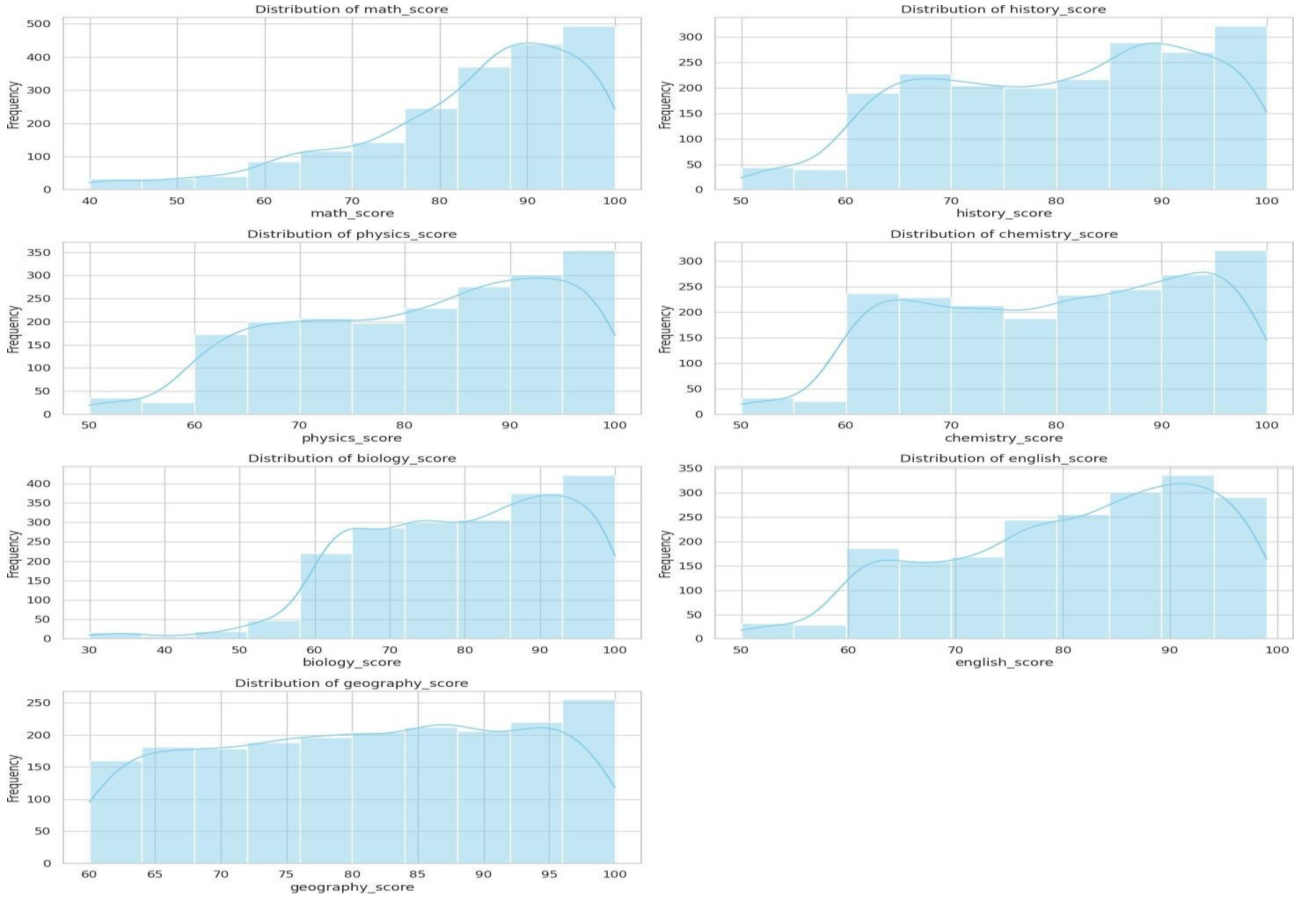

Fig. 2. The Distribution of Subject Scores

In Fig. 2, several histograms have been produced to measure and plot the distribution of students' performance or scores in different subjects. The histograms provide a clear overview of student performance across several subjects, namely mathematics, history, physics, chemistry, biology, English, and geography. In mathematics, physics, and English, the distributions are heavily concentrated within the 80–100 range, indicating that most students performed strongly in these areas with relatively few low scores. Such clustering suggests consistent achievement in these core subjects, reflecting both higher levels of proficiency and possibly greater emphasis in instruction. In contrast, the results for chemistry, biology, and history are more evenly distributed across wider score ranges, reflecting a mixture of average and high performance. While peaks around the 90 marks in history and biology highlight the presence of strong achievers, the overall spread indicates greater variation in performance compared to mathematics and physics.

Geography displays the widest distribution among all subjects, with scores spread across a broad range but showing a modest concentration between 85 and 95. This indicates that while many students performed well, the level of consistency was lower than in other subjects. Taken together, the data highlight two distinct patterns: consistent, high achievement in mathematics, physics, and English, and more varied performance in chemistry, biology, history, and geography. These patterns are significant as they may inform how students' subject strengths influence their academic pathways and career decisions. The results suggest that while students demonstrate solid grounding in technical and language-based disciplines, there remains notable variability in other subject areas, pointing to the need for targeted pedagogical strategies to support more balanced outcomes across the curriculum.

In a bid to produce a career-driven study pathway prediction, a mapping between the career aspirations and the suitable and recommended study pathway is proposed. The suggested mapping can be seen in Table III. Such a guided mapping ensures that students are channeled to academic courses that best suit their dreams and innate capabilities. To keep the options open, the recommended study pathways are either pure science study pathway or social science study pathway.

This study evaluated four hybrid classification models for predicting students' recommended study pathways based on academic profiles and career aspirations (Fig. 3). The baseline ANN (Hybrid A) showed moderate accuracy but was limited by class imbalance and feature interaction issues, consistent with findings by [29], where standalone ANNs performed inconsistently on educational data.

TABLE III. MAPPING OF CAREER ASPIRATIONS TO RECOMMENDED STUDY PATHWAYS

Career Aspiration	**Recommended Study Pathway**
Software Engineer	Pure Science Study Pathway
Business Owner	Social Science Study Pathway
Unknown	Social Science Study Pathway
Banker	Social Science Study Pathway
Lawyer	Social Science Study Pathway
Accountant	Social Science Study Pathway
Doctor	Pure Science Study Pathway
Real Estate Developer	Social Science Study Pathway
Stock Investor	Social Science Study Pathway
Construction Engineer	Pure Science Study Pathway
Artist	Social Science Study Pathway
Game Developer	Pure Science Study Pathway
Government Officer	Social Science Study Pathway
Teacher	Social Science Study Pathway
Designer	Social Science Study Pathway
Scientist	Pure Science Study Pathway
Writer	Social Science Study Pathway

Hybrid B (RF + ANN) improved feature representation by using Random Forest embeddings before ANN classification, yielding better recall and precision. Similar outcomes were observed in [30], which applied Random Forest integrated with Deep Neural Network, and [31], which generally hybridized the models for general classification. However, RF's averaging effect reduced boundary sharpness, limiting fine-grained discrimination. Hybrid C (SVM + ANN) achieved the best overall performance, producing the highest accuracy and balanced F1-score. The SVM layer's margin-based filtering enhanced separability and reduced overfitting—consistent with what observed in [32], which reported superior generalization and interpretability using SVM–ANN frameworks.

Meanwhile, Hybrid D (GBM + ANN) improved minority-class sensitivity but suffered from overfitting. Overall, Hybrid C provided the best trade-off between accuracy, stability, and interpretability, affirming evidence that margin-based neural hybrids outperform other ensemble approaches for educational decision-support [33].

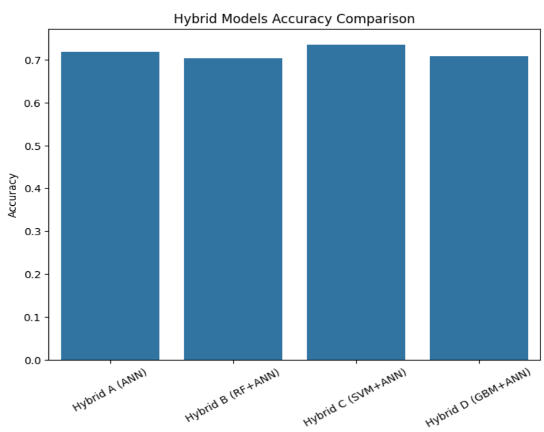

Fig. 3. The Accuracy Achievement Comparison

Fig. 4 shows the confusion matrix recorded for each of the hybrid models trained and tested with the dataset. It can be seen from this figure that the False Negatives recorded in the case of Hybrid A, Hybrid B and Hybrid D are relatively much higher than that of Hybrid C. In other words, the hybrid models mistakenly predict considerable numbers of students, who are supposed to be recommended to enrol for the pure science pathway, to be recommended with the social science pathway.

On another note, it can also be observed from Fig. 5 that Hybrid C, which is integrating Random Forest and Artificial Neural Networks with SMOTE-based imbalanced data handling, exhibits the best accuracy performance by having the highest number of True Positives, recording 23 of them as compared to the other hybrid models which are only capable of recording less than 15 True Positives.

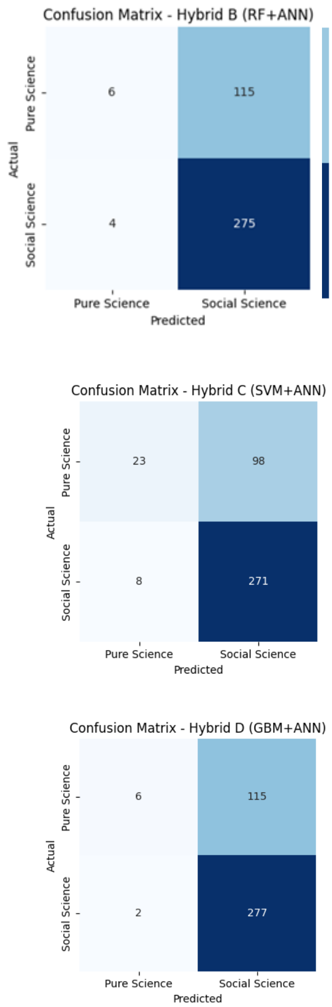

Fig. 4. Confusion Matrix of Hybrid A, B, C and D Models.

VII. CONCLUSION

In conclusion, this study demonstrates the potential of hybrid deep learning models in accurately predicting career-driven study pathways for secondary school leavers. Among the four models evaluated, Hybrid C (SVM+ANN) consistently outperformed the others by achieving the best trade-off between accuracy, stability, and generalization, highlighting the strength of margin-based refinements in enhancing neural network performance. The findings suggest that integrating support vector mechanisms with ANN provides a more robust and interpretable framework for addressing the complexities of educational pathway prediction. This work contributes to the growing body of research on AI-driven educational guidance and offers a promising foundation for developing scalable recommendation approaches to support students in making informed academic and career decisions. Future research should further explore longitudinal validation, fairness considerations, and real-world deployment to maximize the societal impact of such approaches.

Acknowledgement/Funding

The authors would like to acknowledge the support of Universiti Malasyia Kelantan (UMK) for providing the facilities and support on this research.

Conflict of Interest Statement

The authors agree that this research was conducted in the absence of any self-benefits, commercial or financial conflicts and declare the absence of conflicting interests with the funders.

Author's Contributions

Normasliza Malik carried out the research, wrote and revised the article. Hanafi Thman carried out the review. Hadhrami Ab Ghani anchored the review, revisions and approved the article submission.

References

[1] Dirin, A., & Saballe, C. A. (2022). Machine learning models to predict students' study path selection. *International Journal of Interactive Mobile Technologies, 16*(1), 158–183. https://doi.org/10.3991/IJIM.V16I01.20121

[2] Tony, R. (2021). *A study of Malaysian high school leavers' dilemma in choosing varsity courses*. Retrieved from https://www.researchgate.net/publication/352550948_A_Study_of_Malaysian_High_School_Leavers'_Dilemma_in_Choosing_Varsity_Courses

[3] Baker, R. S. J. d., & Siemens, G. (2014). Educational data mining and learning analytics. In R. S. J. d. Baker & G. Siemens (Eds.), *Learning Analytics* (pp. 1–17). Springer.

[4] Ojajuni, O., Ayeni, F., Akodu, O., Ekanoye, F., Adewole, S., Ayo, T., Misra, S., & Mbarika, V. (2021). Predicting student academic performance using machine learning. In *Lecture Notes in Computer Science* (Vol. 12957, pp. 481–491). https://doi.org/10.1007/978-3-030-87013-3_36

[5] Vaidhehi, V., & R., S. (2018). A systematic review of recommender systems in education. *International Journal of Engineering & Technology, 7*(3.4), 188. https://doi.org/10.14419/IJET.V7I3.4.16771

[6] Abdrakhmanov, R., Zhaxanova, A., Karatayeva, M., Niyazova, G. Z., Berkimbayev, K., Tuimebayev, A., Auezov, S. M., Cengiz, M., & Akhmet Yassawi, K. (2024). Development of a framework for predicting students' academic performance in STEM education using machine learning methods. *International Journal of Advanced Computer Science and Applications, 15*(1), 38–46. https://doi.org/10.14569/IJACSA.2024.0150105

[7] Sandeepa, A. G. R., & Mohottala, S. (2025). *Evaluation of machine learning models in student academic performance prediction* (arXiv preprint).

[8] Baashar, Y., Alkawsi, G., Ali, N., Alhussian, H., & Bahbouh, H. T. (2021). Predicting student's performance using machine learning methods: A systematic literature review. In *Proceedings of the International Conference on Communications and Information Systems (ICCOINS)* (pp. 357–362). https://doi.org/10.1109/ICCOINS49721.2021.9497185

[9] Beseiso, M. (2025). Enhancing student success prediction: A comparative analysis of machine learning techniques. *TechTrends, 69*, 372–384.

[10] Nadela, S. D. (2023). *Application of data science in career path and employability analysis from secondary and higher education students: A systematic review* (pp. 1–21).

[11] Kumar, M., Walia, N., Bansal, S., Kumar, G., & Cengiz, K. (2023). Predicting college students' placements based on academic performance using machine learning approaches. *International Journal of Modern Education and Computer Science, 15*(6), 1–15. https://doi.org/10.5815/IJMECS.2023.06.01

[12] Vadivel, B., Alam, S., Nikpoo, I., & Ajanil, B. (2023). The impact of low socioeconomic background on a child's educational achievements. *Education Research International*. https://doi.org/10.1155/2023/6565088

[13] Zhang, F., Jiang, Y., Ming, H., Ren, Y., Wang, L., & Huang, S. (2020). Family socio-economic status and children's academic achievement: The different roles of parental academic involvement and subjective social mobility. *British Journal of Educational Psychology, 90*(3), 561–579. https://doi.org/10.1111/BJEP.12374

[14] Gu, J. (2024). Factors influencing Chinese students' higher education choices. *Interdisciplinary Humanities and Communication Studies, 1*(7). https://doi.org/10.61173/A3B4SY82

[15] Bhardwaj, P., Gupta, P. K., Panwar, H., Siddiqui, M. K., Morales-Menendez, R., & Bhaik, A. (2021). Application of deep learning on student engagement in e-learning environments. *Computers & Electrical Engineering, 93*, 107277. https://doi.org/10.1016/J.COMPELECENG.2021.107277

[16] Alsoqae, A. M. A., & Belal, S. (2024). Student academic success factors in selected nursing colleges in Saudi Arabia. *International Journal of Innovative Research in Medical Science, 9*(6), 317–330. https://doi.org/10.23958/ijirms/vol09-i06/1885

[17] Wang, J., & Yu, Y. (2025). Machine learning approach to student performance prediction of online learning. *PLoS ONE, 20*(1), e0299018.

[18] Cash, T. N., & Oppenheimer, D. M. (2024). Parental rights or parental wrongs: Parents' metacognitive knowledge of the factors that influence their school choice decisions. *PLoS ONE, 19*. https://doi.org/10.1371/JOURNAL.PONE.0301768

[19] Alalawi, K., Athauda, R., & Chiong, R. (2025). An extended learning analytics framework integrating machine learning and pedagogical approaches for student performance prediction and intervention. *International Journal of Artificial Intelligence in Education, 35*, 1239–1287. https://doi.org/10.1007/s40593-024-00429-7

[20] Syed Mustapha, S. M. F. D. (2023). Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods. *Applied System Innovation, 6*(5). https://doi.org/10.3390/asi6050086

[21] Anwer, N., & B., M. (2023, November 27). *Performance prediction of students in higher education using multi-model ensemble approach*. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10328829

[22] So, J. C. H., Ho, Y. H., Wong, A. K. L., Chan, H. C. B., Tsang, K. H. Y., Chan, A. P. L., & Wong, S. C. W. (2023). Analytic study for predictor development on student participation in generic competence development activities based on academic performance. *IEEE Transactions on Learning Technologies, 16*(5), 1–15. https://doi.org/10.1109/TLT.2023.3291310

[23] Chen, J., Zhou, X., Yao, J., & Tang, S. K. (2024). Evaluation of student performance based on learning behavior with random forest model. In *2024 13th International Conference on Educational and Information Technology (ICEIT)* (pp. 266–272). https://doi.org/10.1109/ICEIT61397.2024.10540693

[24] Fagbola, T. M., Adeyanju, I. A., Olaniyan, O., Esan, A., Omodunbi, B., Oloyede, A., & Egbetola, F. (2019). *Development of mobile-interfaced machine learning-based predictive models for improving students performance in programming courses* (arXiv preprint).

[25] Su, Y.-S., Lin, Y.-D., & Liu, T.-Q. (2022). Applying machine learning technologies to explore students' learning features and performance prediction. *Frontiers in Neuroscience, 16*, Article 1018005. https://doi.org/10.3389/fnins.2022.1018005

[26] Dervenis, C., Kyriatzis, V., Stoufis, S., & Fitsilis, P. (2022). Predicting students' performance using machine learning algorithms. In *Proceedings of the 6th International Conference on Algorithms, Computing and Systems (ICACS 2022)*. ACM.

[27] Radhika, R. H. (2016). Application of machine learning algorithms for betterment in education system. In *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*. https://doi.org/10.1109/ICACDOT.2016.7877759

[28] Acharya, A., Sahu, S. K., & Ghosh, S. (2014). Early prediction of students' performance using machine learning techniques. *International Journal of Computer Applications, 107*(1), 37–43. https://doi.org/10.5120/18717-9939

[29] Khairy, D., Alharbi, N., Amasha, M. A., Areed, M. F., & Alkhalaf, S. (2024). Prediction of student exam performance using data mining classification algorithms. *Education and Information Technologies, 29*, 21621–21645. https://doi.org/10.1007/s10639-024-12619-w

[30] Al-Hurdi, A. S. K., & Munassar, N. M. A. (2025). Enhancing students' academic performance classification in e-learning using hybrid model (random forest and deep neural network). *Journal of Science and Technology, 30*(7), 86–98. https://doi.org/10.20428/jst.v30i7.2948

[31] Mazumder, P., & Baruah, S. (2023). A hybrid model for predicting classification datasets based on random forest, support vector machine and artificial neural network. *International Journal of Innovative Technology and Exploring Engineering, 13*(1), 19–25. https://doi.org/10.35940/IJITEE.A9757.1213123

[32] Onyango, L. A., Waititu, A. G., Mageto, T., & Kilai, M. (2022). A hybrid classification model of artificial neural network and non-linear kernel support vector machine. *International Journal of Data Science and Analysis, 8*(2), 47–58. https://doi.org/10.11648/j.ijdsa.20220802.15