INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

OPEN ACCESS

Received: 08-12-2024

Accepted: 13-01-2025

Published: 12-02-2025

A Hierarchical and Multi-Tiered Personalized Career Recommender System Tailored to Individual Aptitudes

M Noushad Rahim[1]*, **K P Mohamed Basheer**[2]

1 Department of Computer Science, Sullamussalam Science College Areekode, University of Calicut, Kerala, India

2 Amal College of Advanced Studies, Nilambur, University of Calicut, Kerala, India

Citation: Rahim MN, Basheer KPM (2025) A Hierarchical and Multi-Tiered Personalized Career Recommender System Tailored to Individual Aptitudes. Indian Journal of Science and Technology 18(3): 231-244. https://doi.org/10.17485/IJST/v18i3.3979

* **Corresponding author**.

mannayilnoushadrahim@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN

Print: 0974-6846

Electronic: 0974-5645

Abstract

Objectives: To design and implement a hierarchical, multi-tiered career recommender system with real-world applications, that addresses data sparsity and cold-start limitations, while tackling the challenges posed by the low student-to-career counselor ratio in a context like India. The system aims to deliver automated, personalized career recommendations tailored to individual aptitudes, effectively guiding students toward appropriate educational pathways and contributing to optimizing the nation's human resource potential. **Methods**: This study has developed a novel hierarchical, multi-tiered recommendation system employing machine-learning models based on a content-based filtering approach. The model is built on a first-of-its-kind dataset comprising 612 unique real-world careers, curated by extracting ability profile information from the Occupational Information Network database (O*NET) and utilized a multi-tiered approach to deliver diverse yet personalized career recommendations, with visualizations of the recommended career clusters against the candidate's user profile, providing clear insights into the rationale behind the system's recommendations. **Findings**: The system surpasses other career recommendation systems across multiple dimensions, demonstrating robust performance with accuracy, precision, recall, and f1-scores all exceeding 99%. It provides diverse career options aligned with the candidate's profile through its unique multi-tiered outputs, while effectively addressing data sparsity and cold-start limitations. **Novelty**: The study curated a unique, first-of-its-kind dataset comprising 612 distinct careers, each mapped with their respective ability profile parameters. A novel hierarchical, multi-tiered recommendation engine was developed, introducing an innovative approach that enables the generation of diverse recommendations, overcoming the limited output capabilities characteristic of existing recommender systems.

Keywords: Artificial Intelligence; Career Counselling; Psychometric Tests; Career Recommender Systems; MultiTiered Architecture; Data Sparsity and Cold-Start Limitations

1 Introduction

Providing the right education, complemented by proper career counselling that considers students' innate abilities, is essential for developing a potential human resource for nation-building. This can ensure that the investment in education is fruitful, mitigating dropouts in high-end institutions and reducing the high rate of job switching in industries. Psychometric-test-based career counselling approaches have been modelled to address these issues by measuring candidates' potential to guide them onto the proper educational track. However, these tools have been criticized for their socio-economic and cultural biases [1], highlighting the need for proper standardization. While advancements in machine learning algorithms have led to the development of career recommender systems based on psychometric assessments, most of these systems face challenges in addressing the aforementioned bias, as well as cold-start and data sparsity limitations [2]. Additionally, many existing career recommender systems are restricted to specific career domains, highlighting the need for a comprehensive system to overcome these limitations.

A Personalized Career-Path Recommender System (PCRS) [3] employs a fuzzy logic-based N-layered architecture and facilitates engineering discipline recommendations based on academic performance, personality types, and extracurricular activities. The academic performance is measured based on major STEM topics, and the personality type is calculated using the Myers-Briggs Type Indicator Test. The advantage of the system is in its consideration of personality type alongside academic performance. However, the system's scope is limited to mere engineering discipline, exhibits cultural bias on regional data, and relies on self-reported data for personality assessment.

An adaptive recommendation system [4] predicts suitable engineering education paths for students after completing their preparatory course by extracting relevant features for each engineering discipline using data mining techniques and models such as KNN, SVM, and NB. While the paper presents a promising recommendation architecture, the recommendations were limited to specific areas like urban planning and mining & petroleum engineering. Additionally, the system was built on an imbalanced dataset, and the feature selection mechanism for different educational domains remains a significant challenge.

A Deep Neural Network recommendation system [5], integrating diverse socio-demographic factors like age, sex, income, grades, and number of siblings along with academic performance, was designed to support career counsellors in guiding students to select their academic trajectories. A substantial cohort comprising 1500 junior high school students in the Philippines was considered in the study, enhancing the validity of the system. The main limitation of the system is its reliance solely on academic and socio-economic variables and its failure to account for the candidate's aptitude, which is essential to quantitatively assess whether they can be successful in the recommended courses and related careers.

A Deep NLP-based Job Recommender System [6] for engineering students across six disciplines was built using a user profile dataset, which includes features such as qualifications, engineering discipline, major subjects, GPA, programming languages, frameworks, and platforms, as well as a job profile dataset that extracts information from related industries. While the NLP-based system can assist engineering students with job recommendations, it is not suitable for recommending career paths for secondary students based on their abilities. Furthermore, it does not tackle issues like data sparsity or cold start problems.

An Early Career Trajectory Recommender System (ECTRS)[7], targeting youth and high school students recommends career paths with a focus on Information Technology pathways. The system exploits the Occupational Information Network (O*NET) database to generate user profiles based on the Holland code tests. Additionally, the study uses questionnaires to identify student skills and knowledge. The career profile was created on the same profile as the user profiles. While this work provides valuable insights for students exploring IT careers, it does not objectively evaluate the user's skills but instead relies on the self-reported skillset and limits recommendations to the IT sector, excluding broader real-world career options.

A content-based filtering (CBF) career recommendation system[8], designed to recommend careers related to computer science, provided personalized career trajectories in the field of computer science. The system provides insights into different careers to the students and compares their career interests with the career information extracted from the JobStreet website using web scraping. The user profiles were compared against career profiles using cosine similarity measures on TF-IDF vectors. The system's recommendations are limited to computer-science-related careers and face challenges like data sparsity issues due to its dependence on web scrapping for career profile generation. Moreover, the system does not measure the skill set of a candidate quantitatively in its recommendation strategy.

A hybrid recommendation system that integrates collaborative filtering (CF) and knowledge-based approaches[9], initially recommends universities and majors by the CF system built on graduate ratings, and these recommendations are further refined by a knowledge-based system built upon Case-Based Reasoning (CBR) and an ontology created from institution, student, and career data. The limitations of the study include its dependency on the survey method of data collection for gathering demographic details, educational histories, interests, and career ratings, leading to challenges such as data bias and quality issues, which may result in a biased system with cold-start limitations. Additionally, the system faces scalability issues.

A multiple intelligence-based career recommender system[10], based on machine learning algorithms like KNN, Decision Trees and XGBoost, recognises the need for a holistic career guidance approach beyond the student's academic performance by considering the unique eight different aspects of human intelligence. However, the study has notable limitations since it depended on the self-reported data which can cause quality issues and related biases in the recommendation process. Furthermore, the system does not address the socio-economic and cultural factors that can influence career decisions.

A Job Recommender System Based on Machine Learning and Ontology[11] developed on machine-learning and ontology-based approaches, supports candidates' employability. The system utilises learners' skills, educational qualifications and work experience. The recommendation system provides suggestions based on the machine-learning model supplemented with an ontology framework for the semantic enrichment of the recommendations. The advantage of the system is the ontology framework provides meaningful recommendations. The system does not address data sparsity and cold start limitations. Additionally, the dataset used is not a comprehensive one to address real-world careers.

1.1 Limitations of Existing Systems and Addressing the Research Gap

The literature review reveals numerous limitations of existing systems, including recommendations limited to a smaller set of careers, thereby restricting candidates' possible career paths; reliance on self-reporting rather than quantitative evaluation of candidates' aptitudes; and failure to effectively address data sparsity and cold-start issues. In a country like India, where the student-to-career counsellor ratio is very low and many career counsellors are not well-versed in psychometric tools, a machine-learning-based career recommendation system addressing bias, data sparsity, and cold-start limitations can significantly improve this situation. The proposed career recommendation system is targeted to address these challenges.

The proposed career recommender system employs Differential Aptitude Tests (DAT)[12] while acknowledging and addressing the criticisms against DAT, to evaluate candidates' inherent abilities and recommend suitable careers. Specifically, we utilised the Kerala Differential Aptitude Tests (KDAT), an aptitude testing tool developed by the Department of General Education, Govt. of Kerala[13]. Considering the region's unique socioeconomic factors, KDAT has been designed and standardised specifically for 10th-grade students in Kerala under the guidance of experts in the relevant fields, minimising potential biases in the testing process (Official Communication, Career Guidance and Adolescent Counselling, Directorate of General Education, November 5, 2024). KDAT, a customised version of the DAT, was designed to offer a comprehensive and standardized framework for assessing the intrinsic abilities of junior and senior high school students. The DAT is built on a wide array of cognitive domains, including cognitive processing, reading comprehension, language proficiency, attention and working memory, executive functioning, processing speed, and complex reasoning and problem-solving abilities. Through a detailed analysis of theoretical constructs eight distinct test batteries were developed[12]. These batteries help to measure a spectrum of aptitudes as shown in Table 1.

Career counsellors typically assess DAT results manually, leading to subjectivity in the counselling process due to the knowledge gap among career counsellors. The survey conducted as part of this research revealed inconsistencies in the interpretation of DAT by career counsellors, as illustrated in Figure 1 for Agricultural Engineering. The scarcity of career counsellors proficient in DAT further underscores the need for a standardised, automated career recommendation system (Official Communication, Career Guidance and Adolescent Counselling, Directorate of General Education, Kerala, November 6, 2024).

The proposed system will provide career recommendations automatically based on the DAT score of a candidate which is generated from KDAT. This study addresses many other issues like cold start and data sparsity issues, low student-to-career counsellor ratio, and subjectivity in career counselling by curating a comprehensive first-of-its-kind dataset containing 612 unique real-world careers complemented with an innovative novel recommender system architecture.

Table 1. Differential Aptitude Test batteries and their interpretations

Ability	Description	Related Area
Numerical Ability (NA)	Proficiency in comprehending numerical relationships and executing arithmetic calculations.	Mathematics, physics, engineerin, etc.
Abstract Reasoning (AR)	Aptitude for recognizing concepts conveyed not through numbers or words but via sequences of figures.	Computer programming, law, economics, engineerin, etc.
Verbal Reasoning (VR)	Ability to conceive ideas framed in words	Law, business, journalism, science, etc.
Mechanical Reasoning (MR)	Proficiency in comprehending fundamental mechanical principles and utilizing tools.	Engineer, mechanic, electrician, carpenter, etc.
Space Relations (SR)	Ability to visualize or think in spatial terms	Architecture, designing, dentistry, civil engineering, etc.
Perceptual Speed and Accuracy(PSA)	Speed and accuracy of response in simple perceptual tasks	Clerical works, filing, coding, etc.
Language Spelling (SP)	Ability to use spelling	Proofreading, law, secretarial work, journalism, etc.
Language Usage (LU)	Ability to use language Grammar	Proofreading, law, secretarial work, journalism, etc.

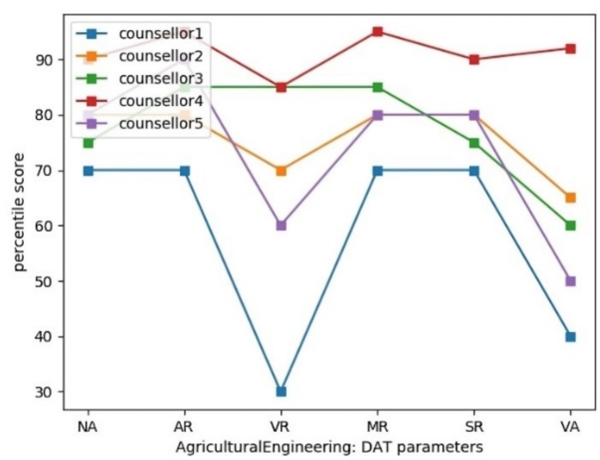

Fig 1. Divergent assessments by career counsellors

2 The Methodology

The proposed framework constitutes a content-based career recommender system based on differential aptitude tests. For the creation of individual user profiles, we administer Differential Aptitude Tests (DAT) to each candidate employing the Kerala Differential Aptitude Tests, a testing tool developed by the Directorate of General Education, Government of Kerala. To derive the career profile, the ability profile information available in the Occupational Information Network database is utilized. This section offers a thorough breakdown of the entire process, including critical stages like data collection for creating career profiles, feature extraction, data augmentation, dataset creation, model building, user profile generation, and other essential components intrinsic to this methodology. The overall architecture of the proposed system is given in Figure 2.

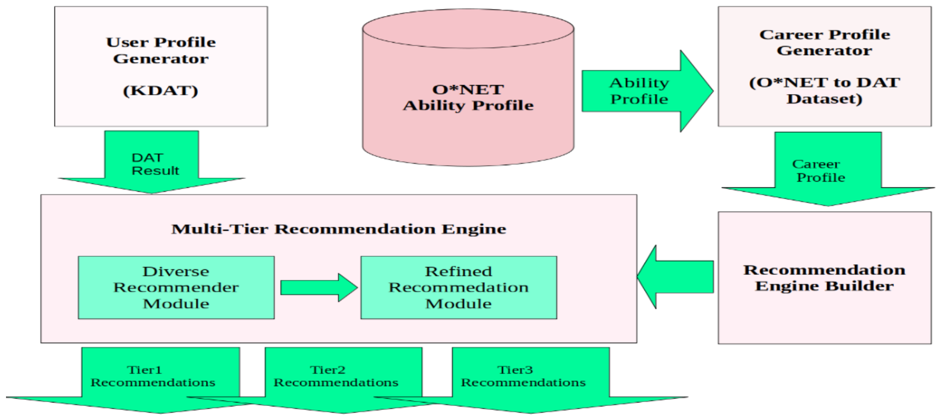

Fig 2. The overall architecture of the proposed recommendation system

2.1 User Profile Generation using KDAT

The user profile is generated by administering DAT for each student utilising the Kerala Differential Aptitude Tests (KDAT). While DAT is made up of a set of eight distinct test batteries as shown in Table 1, KDAT simplifies this structure, blending the Language Spelling (SP) and Language Usage (LU) batteries into a single category termed Verbal Ability. Further, KDAT omits the Perceptual Speed and Accuracy (PSA) battery and consolidates the test into a set of six test batteries: Abstract Reasoning (AR), Numerical Ability (NA), Verbal Reasoning (VR), Mechanical Reasoning (MR), Space Relations (SR), and Verbal Ability. The 180 unique questions in KDAT are strategically distributed across the above-mentioned six aptitude categories. These questions are instrumental in assessing the skill sets of candidates in their respective domains. The calculation of percentile scores for candidates is done through a process utilizing parameters like difficulty index, discrimination factor, and guess factor associated with each question.

2.1.1 Normalization and Standardisation of User Profile Data

Following the administration of the DAT, the user profile is acquired as represented below:

$DAT_{userprofile}$ = [X_1, X_2, X_3, X_4, X_5, X_6]

Where each X_i represents the DAT score for each test battery.

Since each test battery score of DAT is on a percentile scale ranging from 0 to 100, normalization is not applied to the user profile data. However, standardization is employed on the DAT scores using the formula as shown in Equation (1).

$$Z_i = \frac{(X_i - \mu(X))}{\sigma(X)} \tag{1}$$

Where,

X_i is the percentile score obtained by the candidate for each test battery,

$\mu(X)$ is the average of scores across all test batteries,

$\sigma(X)$ is the standard deviation of scores across all test batteries, and

Z_i is the standardized value for the ith test battery.

The final user profile record can be represented as:

Standardized $DAT_{userprofile}$ = [Z_1, Z_2, Z_3, Z_4, Z_5, Z_6]

2.2 Career Profile Generation

To construct a robust career recommender system grounded in differential aptitude tests, a comprehensive dataset encompassing all feasible real-world careers, correlated with their requisite skill sets derived from DAT parameter values, is fundamental. Unfortunately, despite the existence of multiple career-related datasets, a publicly accessible dataset based on DAT for generating career profiles has not been available to date. To overcome this limitation, we devised a systematic methodology for generating a dataset tailored to career profile development. Our approach involves harnessing the ability profile data within the Occupational Information Network (O*NET) database[14]. This resource provides a wealth of information regarding abilities, skills, tasks, and other pertinent attributes pertinent to diverse careers.

2.2.1 Feature Engineering and Dataset Creation

Using the ability profile data in the O*NET, we orchestrated the creation of a dataset by meticulously mapping the substantial features associated with over 612 distinct careers extracted from O*NET to the six DAT test batteries. For each career entry in O*NET, comprising 52 unique features, we discerned and mapped the most consequential ones to our specific DAT test batteries. The mapping of the O*NET ability profile to DAT test batteries is given in Table 2. The process of mapping is guided and informed by expert domain knowledge in DAT.

To convert the ability profile parameter values into their corresponding DAT test battery values, we utilized the Importance and Level scale values from the O*NET ability profile parameters. Normalization was necessary due to the differing scales: Importance values ranged from 1 to 5, and Level values ranged from 0 to 7. The normalization process is done, employing the standard normalization equation, as shown in Equation (2).

$$X_{normalized} = \frac{X - min(X)}{max(X) - min(X)} \tag{2}$$

Where X represents the original *Importance* score or *Level* score.

Table 2. Mapping of DAT test batteries to O*NET ability profile parameters

DAT Parameter	Ability profile of Occupational Information Network Database
Numerical Ability	Mathematical Reasoning, Number Facility
Abstract Reasoning	Inductive Reasoning, Deductive Reasoning, Fluency of Ideas
Verbal Reasoning	Oral Comprehension, Written Comprehension
Mechanical Reasoning	Problem Sensitivity, Manual Dexterity, Finger Dexterity, Control Precision
Space Relations	Visualization
Verbal Ability	Written Expression
Clerical Speed and Accuracy	Perceptual Speed

Given that Importance values fall between 1 to 5 and Level values span from 0 to 7, we computed normalized scores using Equation (3) and Equation (4).

$$Score_{Importance} = \frac{I_i - 1.0}{5.0 - 1.0} \tag{3}$$

$$Score_{Level} = \frac{L_i - 0.0}{7.0 - 0.0} \tag{4}$$

Where I_i and L_i represent the Importance score Level score for parameter i respectively.

After normalizing the values for each mapped ability profile parameter, we employed Equation (5) to calculate the corresponding DAT parameter values in percentile.

$$X_i = \frac{Score_{Importance} + Score_{Level}}{2.0} * 100 \tag{5}$$

Where X_i is the score in the percentile of the i^{th} DAT test battery.

Table 3 outlines the process of generating a career profile for the profession ‘Lawyer’ using the aforementioned procedure. Hence, the skill set for the career Lawyer in terms of DAT parameters is determined as:

DAT_{Lawyer}*(NA, AR, VR, MR, SR, VA)* $_=$ *[37, 67, 80, 30, 37, 75]*

This approach was used to calculate the DAT parameter values for 612 distinct careers, establishing the foundational dataset for our career recommendations.

Table 3. The process of generating career profile for the profession *Lawyer*

DAT Test Battery	Mapped O*NET Ability Parameters	Importance Score	Level Score	Conversion Process	Computed Value (Rounded Up)	Finding Average Value	Final DAT Score
Numerical Ability	Mathematical Reasoning	2.62	2.38	(((2.62-1)/(5-1))*100+((2.38-0)/(7-0))*100)/2	38	(38+36)/2	37
	Numerical Facility	2.38	2.5	(((2.38-1)/(5-1))*100+((2.5-0)/(7-0))*100)/2	36		
Abstract Reasoning	Inductive Reasoning	4	4.25	(((4.0-1)/(5-1))*100+((4.25-0)/(7-0))*100)/2	68	(68+69+63)/3	67
	Deductive Reasoning	4	4.38	(((4.0-1)/(5-1))*100+((4.38-0)/(7-0))*100)/2	69		
	Fluency of Ideas	3.75	3.88	(((3.75-1)/(5-1))*100+((3.88-0)/(7-0))*100)/2	63		
Verbal Reasoning	Oral Comprehension	4.5	5	(((4.5-1)/(5-1))*100+((5.0-0)/(7-0))*100)/2	80	(80+80)/2	80
	Written Comprehension	4.5	5	(((4.5-1)/(5-1))*100+((5.0-0)/(7-0))*100)/2	80		
Mechanical Reasoning	Problem Sensitivity	3.88	4	(((3.88-1)/(5-1))*100+((4.0-0)/(7-0))*100)/2	65	(65+8+30+15)/4	30
	Manual Dexterity	1.38	0.38	(((1.38-1)/(5-1))*100+((0.38-0)/(7-0))*100)/2	8		
	Finger Dexterity	2.25	2	(((2.25-1)/(5-1))*100+((2.0-0)/(7-0))*100)/2	30		
	Control Precision	1.75	0.75	(((1.75-1)/(5-1))*100+((0.75-0)/(7-0))*100)/2	15		
Space Relations	Visualization	2.5	2.5	(((2.5-1)/(5-1))*100+((2.5-0)/(7-0))*100)/2	37	37	37
Verbal Ability	Written Expression	4.12	5	(((4.12-1)/(5-1))*100+((5.0-0)/(7-0))*100)/2	75	75	75

2.2.2 Generating Career Clusters

After the generation of the dataset comprising 612 distinct occupational roles, a clustering methodology is applied to identify career clusters characterized by shared ability profiles. Various clustering algorithms, including Agglomerative Clustering, Gaussian Mixture, Birch, and K-Means, were employed and the ultimate choice of the K-Means clustering approach was determined based on the Silhouette Scores[15] for each cluster. Finally, the optimal number of clusters was determined by applying the silhouette score (Equation (6)) together with the elbow method (Equation (7))[16].

Silhouette Score,

$$S(k) = \frac{1}{N} \sum_{N=1}^{N} \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{6}$$

Where,

S(k) is the silhouette score for a specific number of clusters (k),

N represents the total number of data points,

a(i) denotes the average distance from the i^{th} data point to other data points within the same cluster,

b(i) signifies the smallest average distance from the i^{th} data point to data points in a different cluster, minimized over clusters.

K-means,

$$WCSS_k = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(X_{ij} - \mu_i\right)^2 \tag{7}$$

Where,

$WCSS_k$ is the Within-Cluster Sum of Squares for a specific cluster count k,

n signifies the total number of data points,

X_{ij} represents the jth feature of the ith data point,

μ_i is the centroid of the cluster.

The visual analysis of the eight clusters generated using the K-means algorithm revealed significant variances in attribute values, suggesting the need for further stratification. Consequently, these eight clusters were further stratified into 49 second-level clusters. Figure 3 illustrates a sample of the generated clusters, presenting them alongside their respective DAT parameter values. Thus, the structure consists of eight top-level clusters and 49 lower-level subclusters.

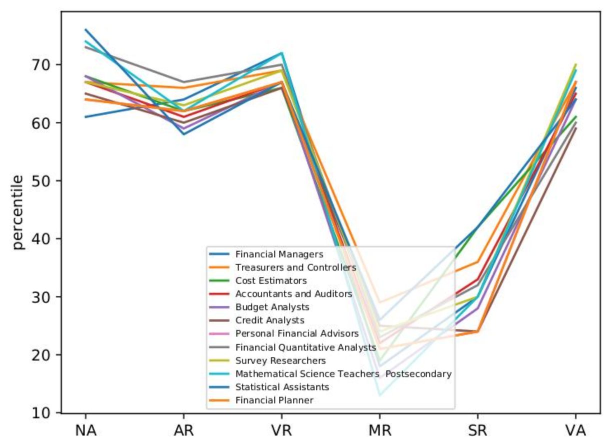

Fig 3. Samples from the generated cluster

The recommendation process adopts a semi-supervised classification approach by manually labelling each record in the generated clusters to build ML-based classifiers. These models consist of a diverse range of algorithms, including ANN, Random Forest Classifier (RF), SVM, KNN, Decision Tree, and Gaussian Naive Bayes (GNB), forming the foundation of the proposed Career Recommender System.

Different ML-based classifiers were tried on the dataset and the data presented in Table 4 highlights the risk of over-fitting [(17)], primarily due to the substantial differences between training and testing accuracies. This difference can be attributed to the dataset's limited size, containing just 612 records.

Table 4. Training and testing accuracies of various models on the original dataset

Sl No	Classifier	Training Accuracy	Testing Accuracy
1	ANN	1.00	0.72
2	Random Forests Classifier	1.00	0.75
3	Decision Tree	1.00	0.63
4	GNB	0.95	0.87
5	KNN	0.86	0.70
6	SVM	0.95	0.87

To address the challenges posed by overfitting, we took measures to augment the records within each cluster of the dataset, which resulted in the generation of 57,109 additional records. The following section provides a comprehensive explanation of the employed augmentation process.

2.2.3 Data Augmentation Procedure

The overfitting problem arises from the dataset's limited size, comprising 612 records distributed across 44 distinct clusters which hinders the classifiers' ability to generalize patterns within these clusters. To address this issue, augmenting each record related to a career within the dataset is essential, enhancing the model's generalization ability [(18)]. We deliberately selected

Gaussian Noise data augmentation [19] to generate extra related records to introduce random fluctuations into the dataset. This approach fosters resilience to noise, promotes the learning of more generalized features, injects diversity into the dataset through random disturbances and serves as a regularization technique. We applied Gaussian Noise data augmentation within our dataset using the mathematical expression given in Equation (8).

$$Y_i = X_i + \varepsilon_i \tag{8}$$

Where,

Y_i represents i^{th} element of the augmented data point,

X_i represents ith element of the original data point,

ε_i is a random value drawn from a Gaussian distribution with a mean of 0

Algorithm 1 outlines the entire process. For our specific augmentation procedure, we drew random values from a Gaussian distribution with a mean(μ) of -2.0 and a noise standard deviation (σ) of 5.0. This process was iterated for each element of the data point with an augmentation factor (F) of 100, resulting in a dataset enriched with controlled noise.

Algorithm 1 Data Augmentation Algorithm

Require: Input data point X_i, Augmentation Factor F, Noise standard deviation σ

Ensure: Augmented data list ADL

for k $\leftarrow$ 1 to F **do**

  for i $\leftarrow$ 1 to length(X_i) **do**

    Generate Gaussian noise for each data point i:

    $\varepsilon_i \leftarrow$ random sample from a Gaussian distribution with mean -2.0 and standard deviation σ = 5.0

    Generate Augmented Record:

      $Y_i \leftarrow X_i + \varepsilon_i$

    Append the record Y_i to the Augmented data list ADL

  end for

end for

return Augmented data list ADL

By applying algorithm 1, we augmented each record within every cluster of the original dataset, consisting of 612 records. As a result, our dataset has been expanded to include 57,109 records. Figure 4 illustrates an original cluster and its related augmented cluster.

2.2.4 Evaluation of Augmented Dataset Validity

To validate the augmented dataset, we built various machine-learning models using the augmented data while retaining the original dataset as the holdout.Table 5 presents the models' performance on the train-test split of the augmented dataset, demonstrating that the overfitting issue has been effectively resolved.

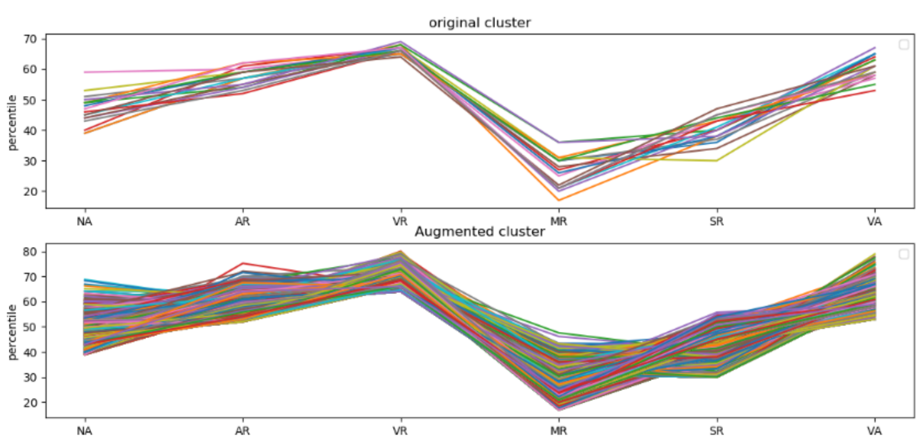

Fig 4. A sample from the augmented clusters

Furthermore, since 57,109 augmented records were generated from just 612 records in the original dataset, it is essential to evaluate whether the augmented dataset captures the patterns of the original dataset. To achieve this, the model built on the augmented dataset was tested using the original dataset records, and the system provided 100% accuracy, confirming that the augmented dataset faithfully captures the patterns present in the original dataset.

Table 5. Performance Evaluation on Augmented Dataset: Training and Testing Accuracies of Various Algorithms

Sl No	Classifier	Training Accuracy	Testing Accuracy
1	ANN	0.97	0.94
2	Random Forests Classifier	1.00	0.99
3	KNN	0.98	0.97

2.3 Model Building

The proposed system employs a novel hierarchical recommendation approach to provide diverse career recommendations tailored to a candidate's DAT results. As previously discussed, the original dataset underwent an initial division into eight distinct clusters based on insights derived from the silhouette and elbow methods. Subsequently, these clusters were further subdivided into 49 clusters, labelled, and augmented to yield the final dataset. Thus, we have a higher-level dataset of eight distinct career clusters and a fine-tuned set of 49 lower-level career clusters. We utilized this hierarchy of career clusters to develop a hierarchical career recommender system.

The recommendation system consists of a Diverse Recommender Module followed by a Refined Recommendation Module. The Diverse Recommendation Module tries to provide diversity in career recommendations at a higher level. The Refined Recommendation Module fine-tunes each set of recommendations delivered by the Diverse Recommendation Module.

The Diverse Recommendation Module is constructed using the initially generated dataset containing top-level eight career clusters. Simultaneously, the Refined Recommendation Module, a set of eight second-level classifiers was developed on datasets generated by further sub-clustering the original top-level career clusters. These second-level refinement classifier models named from RefMod 1 to RefMod 8, act as a Refinement Suite to fine-tune the recommendations made by the top-level Diverse Recommendation Module. Both Diverse Recommendation Module and Refined Recommendation Module are modelled employing voting classifiers, incorporating machine learning algorithms such as ANN, KNN, SVM, Random Forest Classifier, Decision Tree, and Gaussian Naive Bayes Classifier.

The candidate's DAT results serve as input for the Diverse Recommendation Module. This classifier module identifies the best three diverse career classes from the eight available in the model. These three diverse classes are fed to the Refined Recommendation Module. The module then loads the corresponding three second-level classifiers from the Refinement Suite containing eight different classifiers to fine-tune the recommendations made by the Diverse Recommendation Module. Each of these three sub-models, in turn, provides the best three sets of career classes called First Tier, Second Tier and Third Tier career recommendations in the order of their significance for the candidate, resulting in a total of nine different sets of careers. This comprehensive array of options provides the candidate with diverse career paths aligned with their DAT results. Additionally, career counsellors can derive valuable insights from the recommender system to enhance career counselling sessions. Algorithm 2 offers a clearer explanation, while Figure 5 visually illustrates the entire process.

Algorithm 2 The Hierarchical Recommendation Process

```
user_profile ← GetUserProfile()   // DAT result for the candidate
top_classes ← GetBest3Classes(DiverseRecommenderModule, user_profile)
for each class in top_classes do
    refinedModel ← LoadRefinedModel(RefinementSuite, class)        // Obtain predictions from refinedModel
    tier1_predictions ← GetTier1Predictions(refinedModel, user_profile)
    tier2_predictions ← GetTier2Predictions(refinedModel, user_profile)
    tier3_predictions ← GetTier3Predictions(refinedModel, user_profile)
end for
```

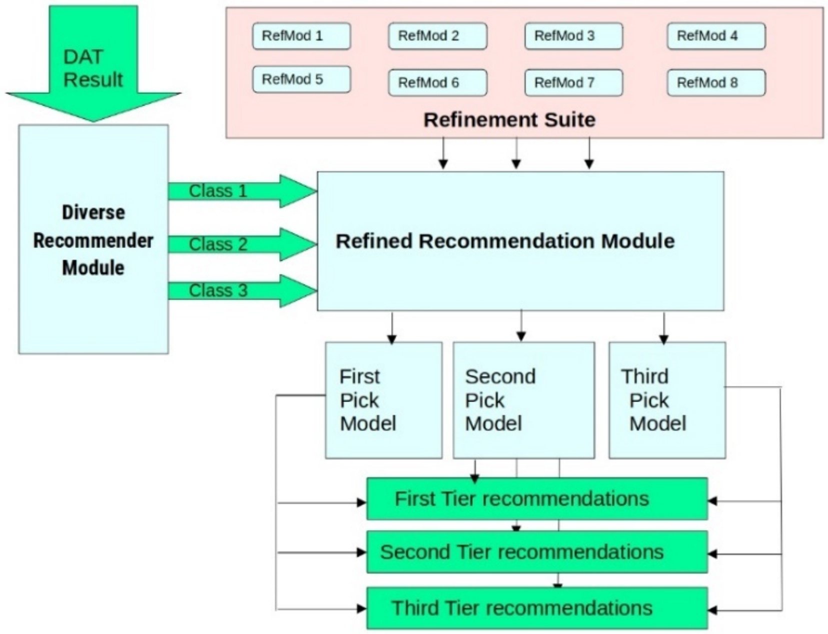

Fig 5. The architecture of the proposed hierarchical recommendation engine

3 Results and Discussions

The system has demonstrated exemplary performance, showcasing testing accuracy exceeding 99% for both the top-level Diverse Recommender Module and the sub-level Refined Recommendation Module. In addition to this, crucial performance metrics such as precision, recall, and F1 score also surpass 99%, exhibiting highly congruent values. This remarkable consistency underscores the model's robust and dependable predictive capabilities across a spectrum of performance dimensions. In Table 6, we present the testing accuracy results for both the top-level Diverse Recommender Module and the sub-level Refined Recommendation Module, shedding light on the performance of each component.

The model produces varied career recommendations for a candidate by inputting their DAT results into the system. It further refines these recommendations through sub-models. Each sub-model offers multi-tiered recommendations ranked by their significance.Figures 6 and 7 and Figure 8 illustrate recommendations categorised as First Tier, Second Tier, and Third Tier, respectively. These recommendations are generated by the system utilising its First Pick, Second Pick, and Third Pick fine-tuned models based on the user profile:

$DAT_{userprofile}$*(NA, AR, VR, MR, SR, VA) = [37, 67, 80, 30, 37, 75]*

Table 6. Metrics for the Diverse Recommender Module and Refinement Recommendation Modules in the proposed model

Models	Accuracy	Precision	Recall	F1 Score
Diverse Recommender	0.99220	0.99223	0.99221	0.99220
RefMod1	0.99468	0.99472	0.99468	0.99468
RefMod2	0.99729	0.99729	0.99729	0.99729
RefMod3	0.99615	0.99618	0.99615	0.99615
RefMod4	0.99511	0.99512	0.99511	0.99510
RefMod5	0.99703	0.99704	0.99703	0.99703
RefMod6	0.99292	0.99297	0.99292	0.99292
RefMod7	0.99488	0.99491	0.99488	0.99487
RefMod8	0.99434	0.99436	0.99434	0.99434

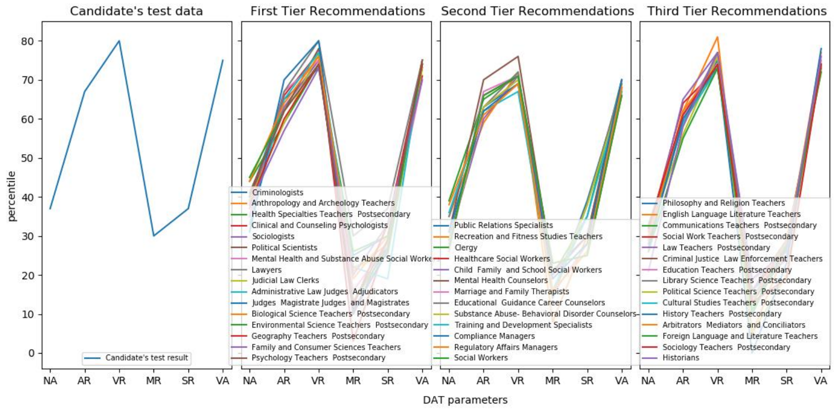

Fig 6. Career recommendations from the FirstPick Model

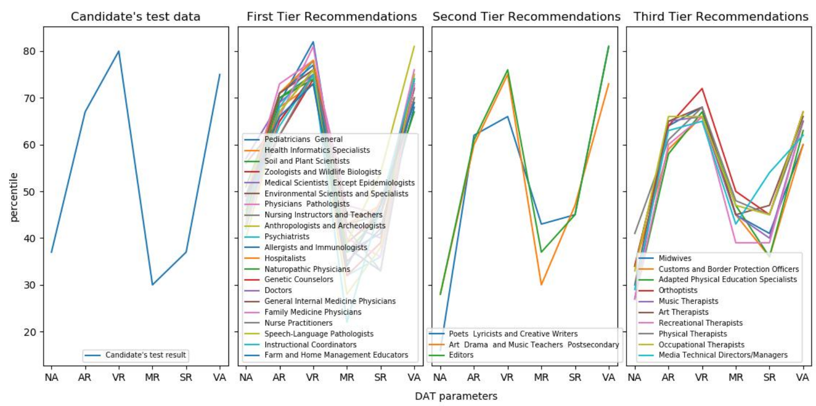

Fig 7. Career recommendations from the SecondPick Model

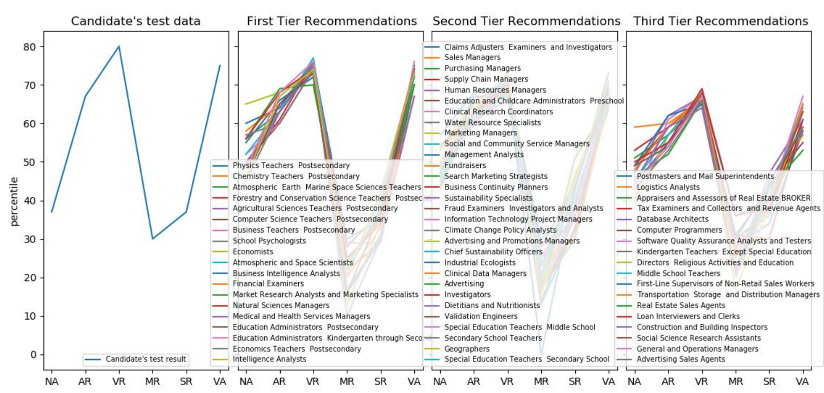

Fig 8. Career recommendations from the ThirdPick Model

3.1 Distinctive Contributions of the Proposed System: A Comparative Study

This study created a first-of-its-kind dataset containing 612 unique careers and their related aptitude parameter values. Additionally, this study introduces a novel hierarchical multi-tiered recommendation architecture to provide diversity in the recommendation process which can be adapted in recommender systems in different domains. The proposed system demonstrates superior performance across multiple dimensions against existing systems. A pivotal feature that distinguishes our system is its ability to recommend almost all conceivable careers in the real world, in contrast to numerous existing systems that restrict recommendations to specific domains, such as computer science-related courses [3,8] or broader academic streams like science, commerce, and humanities without specifying actual careers [5], as well as those exclusively focused on engineering-related [3,6] or information technology-related careers [7]. While the existing system can recommend a limited set of careers, the proposed system, with its novel hierarchical multi-tiered architecture, can offer a diverse array of nine sets of careers from its three output tiers, enriching the user experience. This added advantage, along with proper user profile creation using KDAT, addresses data sparsity and cold start issues, outperforming existing systems [6,9,11]. Using visualization techniques, the system also provides transparency for its recommendations, which is highly advocated today [20,21]. Furthermore, the system excels in accuracy, boasting an impressive 99% accuracy rate, surpassing the 83% to 95% range observed in the reviewed papers.

3.2 Mitigating Cold Start and Data Sparsity Constraints

Every new user undergoes a Differential Aptitude Test consisting of 180 unique questions across various aptitude domains to identify their innate abilities within the proposed career recommendation framework. This allows the system to create a user profile that addresses cold-start issues. Additionally, the system utilizes a comprehensive dataset of 612 real-world careers from O*NET profiles, each defined by aptitude parameters, to recommend careers based on the candidates' abilities. The novel hierarchical, multi-tiered architecture, incorporating content-based filtering, enables diverse, aptitude-aligned career recommendations, effectively addressing data sparsity and item cold-start challenges.

4 Conclusion

The proposed system offers an effective career recommendation tool for secondary school students based on their innate abilities. Key contributions include the creation of a unique dataset of 612 real-world careers and a novel hierarchical, multi-tiered recommendation architecture. The system addresses challenges like data sparsity and cold-start issues while showcasing diverse careers aligned with candidates' ability profiles. It reduces subjectivity in career counselling by providing standardized recommendations. This system can help tackle the low student-to-career counsellor ratio in countries like India, reaching every student statewide. Additionally, it enhances transparency by visualizing career clusters against candidates' profiles. With a testing accuracy of 99%, and precision, recall, and F1-score exceeding 99%, the system outperforms existing solutions.

The work can be enhanced by incorporating additional dimensions like personality traits and vocational interests of a candidate to provide him with a more holistic recommendation. Further study may lead to integrating an eXplainable Artificial Intelligence (XAI) framework into the proposed system to make it a more responsible and unbiased career recommendation system.

Authors' contribution

Noushad Rahim M: Conceptualization, Writing – original draft, Writing – review & editing, Analysis, Supervision. **Dr. Mohamed Basheer K .P:** Conceptualization, Writing – review & editing, original draft, Supervision.

Ethics approval

This study was approved by the ethics committee of Sullamussalam Science College, Areekode, Affiliated with the University of Calicut, Kerala, India.

Consent to participate

Informed consent was obtained from all individual participants included in the study.

References

1) Rynolds CR, Altmann RA, Allen DN. The problem of bias in psychological assessment. In: Mastering Modern Psychological Testing. Springer International Publishing. 2021;p. 573–613. Available from: https://link.springer.com/chapter/10.1007/978-3-030-59455-8_15.
2) Karabila I, Darraz N, El-Ansari A, Alami N, Mallahi M. Addressing data sparsity and cold-start challenges in recommender systems using advanced deep learning and self-supervised learning techniques. *Journal of Experimental & Theoretical Artificial Intelligence*. 2024;p. 1–31. Available from: https://www.tandfonline.com/doi/full/10.1080/0952813X.2024.2401364.
3) Qamhieh M, Sammaneh H, Demaidi MN. PCRS: personalized career-path recommender system for engineering students. *IEEE Access*. 2020;8:214039–214049. Available from: https://ieeexplore.ieee.org/document/9268112.
4) Ezz M, Elshenawy A. Adaptive recommendation system using machine learning algorithms for predicting student's best academic program. *Education and Information Technologies*. 2020;25:2733–2746. Available from: https://doi.org/10.1007/s10639-019-10049-7.
5) Atienza JRD, Hernandez RM, Castillo RL, De Jesus NM, Buenas LJE. A Deep Neural Network in a Web-based Career Track Recommender System for Lower Secondary Education. In: and others, editor. 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), 26-28 August 2022, Ravet, India. IEEE. 2022;p. 1–6. Available from: https://ieeexplore.ieee.org/document/9908965.
6) Saeed T, Sufian M, Ali M, Rehman AU. Convolutional Neural Network Based Career Recommender System for Pakistani Engineering Students. In: and others, editor. 2021 International Conference on Innovative Computing (ICIC), 09-10 November 2021, Lahore, Pakistan. IEEE. 2021;p. 1–10. Available from: https://ieeexplore.ieee.org/document/9715788.
7) Ngiam ATC, Islam MR, Xu G. ECTRS: A Personalised Early Career Trajectory Recommender System for Youths. In: and others, editor. 2022 9th International Conference on Behavioural and Social Computing (BESC), 29-31 October 2022, Matsuyama, Japan. IEEE. 2022;p. 1–7. Available from: https://ieeexplore.ieee.org/document/9995605.
8) Rashid AHA, Mohamad M, Masrom S, Selamat A. Student Career Recommendation System Using Content-Based Filtering Method. In: and others, editor. 2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS), 07-08 September 2022, IPOH, Malaysia. IEEE. 2022;p. 60–65. Available from: https://ieeexplore.ieee.org/document/9918766.
9) Obeid C, Lahoud C, Khoury HE, Champin PA. A novel hybrid recommender system approach for student academic advising named COHRS, is supported by case-based reasoning and ontology. *Computer Science and Information Systems*. 2022;19(2):979–1005. Available from: http://elib.mi.sanu.ac.rs/files/journals/csis/57/csisn57p979-1005.pdf.
10) Mejia MS, Campillo C, Martínez-Santos JC. Career recommendation system for validation of multiple intelligence to high school students. In: Applied Computer Sciences in Engineering. Springer International Publishing. ;p. 110–120. Available from: https://www.researchgate.net/publication/354912875_Career_Recommendation_System_for_Validation_of_Multiple_Intelligence_to_High_School_Students.
11) Dascălu MI, Hang A, Puskás IF, Bodea CN. CareProfSys: a job recommender system based on machine learning and ontology to support learners' employability at regional level. *Issues in Information Systems*. 2023;24(3):71–82. Available from: https://iacis.org/iis/2023/3_iis_2023_71-82.pdf.
12) Setiawati FA. Aptitude Test's Predictive Ability for Academic Success in Psychology Student. *Psychological research and intervention*. 2020;3(1):1–12. Available from: https://www.researchgate.net/publication/345968488_Aptitude_Test's_Predictive_Ability_for_Academic_Success_in_Psychology_Student.
13) Directorate of Higher Secondary Education Thiruvananthapuram. Kerala Higher Secondary Education Department. 2024. Available from: https://www.careerguidance.dhse.kerala.gov.in.
14) The U S Department of Labor, Employment and Training Administration (USDOL/ETA). Browse by Abilities. . Available from: https://www.onetonline.org/find/descriptor/browse/1.A.
15) Shahapure KR, Nicholas C. Cluster quality analysis using silhouette score. In: and others, editor. 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 06-09 October 2020, Sydney, NSW, Australia. IEEE. 2020;p. 747–748. Available from: https://ieeexplore.ieee.org/document/9260048.
16) Saputra DM, Saputra D, Oswari LD, Advances in Intelligent Systems Research. Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method. In: and others, editor. Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019);vol. 172. Atlantis Press. 2020;p. 341–346. Available from: https://www.researchgate.net/publication/341389979_Effect_of_Distance_Metrics_in_Determining_K-Value_in_K-Means_Clustering_Using_Elbow_and_Silhouette_Method.
17) Bejani MM, Ghatee M. A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review*. 2021;54:6391–6438. Available from: https://link.springer.com/article/10.1007/s10462-021-09975-1.
18) Raileanu R, Goldstein M, Yarats D, Kostrikov I, Fergus R. Automatic data augmentation for generalization in deep reinforcement learning. *arXiv*. 2020. Available from: https://arxiv.org/abs/2006.12862.
19) Lashgari E, Liang D, Maoz U. Data augmentation for deep-learning-based electroencephalography. *Journal of Neuroscience Methods*. 2020;346. Available from: https://doi.org/10.1016/j.jneumeth.2020.108885.
20) Ehsan U, Liao QV, Muller M, Riedl MO, Weisz JD. Expanding explainability: Towards social transparency in AI systems. *Proceedings of the 2021 CHI conference on human factors in computing systems*. 2021;p. 1–19. Available from: https://doi.org/10.1145/3411764.3445188.
21) Yang W, Wei Y, Wei H, Chen Y, Huang G, Li X, et al. Survey on explainable AI: From approaches, limitations and applications aspects. *Human-Centric Intelligent Systems*. 2023;3:161–188. Available from: https://link.springer.com/article/10.1007/s44230-023-00038-y.