International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

IJISAE

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Random Forest Model for Prediction of Software Engineering Skill Set among Computer Science Students through Explainable AI

Jasmin Nizar [1], R. Sharmila[2], K. U. Jaseena[3]

Submitted: 06/02/2024 **Revised**: 14/03/2024 **Accepted**: 20/03/2024

Abstract*:* Student skill evaluation is an essential part of education as it gives information on each student's unique talents, strengths, abilities, and areas for development. The purpose of skill-based education in software engineering is to close the knowledge gap between courses of study and industry demands, so graduates can make valuable contributions in a professional software development setting. This method emphasises the value of actual skills in addition to academic knowledge which is in line with the dynamic and quick evolving nature of the software business. A well-rounded set of Soft skills, Life skills, and Technical skills is frequently cited as the reason for the success of individuals working on software development projects. In today's educational landscape, predicting students' skill sets is imperative, encompassing a spectrum of capabilities ranging from Soft and Life skills to Technical expertise. Achieving equilibrium among these proficiencies is crucial for excelling in the ever-changing and cooperative milieu of software development endeavors. This research introduces a novel predictive framework leveraging Random Forest (RF) Algorithm, Principal Component Analysis (PCA) and Explainable Artificial Intelligence (XAI) for software engineering students skillset prediction. The purpose of Random Forest in skillset prediction is to enhance predictive accuracy and robustness by aggregating the outputs of multiple decision trees. To further optimize the efficiency of the proposed model, this study incorporates Principal Component Analysis that ensures the extraction of high-quality and relevant features. Additionally, the study employs Explainable AI techniques using SHAP to identify key features crucial for accurate predictions. The performance of the proposed classification model is evaluated using metrics like accuracy, precision, recall, F1 score, and the Area Under Curve (AUC) value. The simulation results indicate that the recommended PCA-enhanced Random Forest using the XAI model exhibits superior predictive accuracy compared to the baseline machine learning models.

Keywords*—Skillset, Software Engineering, Explainable artificial intelligence, Principal Component Analysis, Random Forest, Machine Learning, SHAP.*

1. INTRODUCTION

In our interconnected modern world, the pervasive influence of software is evident across various aspects of our lives, making it a standout innovation in the contemporary technological era. Operating as a significant, inspiring, and intricate force, software transcends industries and reshapes our communication, work processes, and interactions. Serving as the bedrock of technological progress, software plays a crucial role in propelling innovation, fostering connectivity, and orchestrating the overarching transformation of our digital landscape. Within the dynamic and ever-evolving dominion of the software industry, skill development emerges as a fundamental element for success. In an era marked by rapid technological progress, professionals must continuously refine their skills to remain relevant and effective. The significance of skill development in the software industry cannot be overstated, as it not only empowers individuals with the technical acumen required to adapt to emerging technologies but also cultivates a culture of innovation and creativity. Given the industry's unrelenting pace of change, professionals who prioritise skill development gain a competitive edge, ensuring them to be remain highly sought-after assets for employers.

Despite the omnipresence of the software industry, it grapples with a high rate of project failures, revealing a notable gap between industry demands and the preparedness of software engineering graduates. The assertion that graduates lack the requisite skills for the dynamic software industry raises concerns about the efficacy of current educational approaches. There appears to be a deficiency in both the practical skills needed for software engineering tasks and the inclusion of experiential learning in educational methodologies. This gap between industry expectations and educational outcomes poses a potential barrier between the Information Technology (IT) sector and the Education System.

Experts concern about relying solely on students' self-learning abilities and advocate for a proactive approach to skill development from the onset of their college journey. This acknowledgment underscores the reality while achieving professional success is improbable without honing essential skills, whether navigating technical complexities,

[1]*Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, Tamilnadu, India ,*[0000-0001-9904-959X]*
[2] *Professor, Department of Computer Applications, Karpagam Academy of Higher Education, Coimbatore, Tamilnadu, India [0009-0002-8667-0605]*
[3] *Assistant Professor, Department of Computer Applications, MES College Marampally, Aluva, Kochi, Kerala, India [0000-0003-2315-9578]*
e-mail: jasminnizar.nizar@gmail.com, sharmi.saaravan0521@gmail.com, jaseena.mes@gmail.com

mastering interpersonal dynamics, or cultivating life skills. The deliberate cultivation of these competencies is indispensable for a well-rounded and prosperous career trajectory. As software engineers embark on their professional journeys, challenges often arise due to a misalignment between the skills acquired during university education and workplace expectations [1]. This discrepancy necessitates significant industry investments to train personnel who may not be adequately prepared for various roles within software engineering. The gap between educational curricula and industry demands highlights the need for a more cohesive and tailored approach for preparing the aspiring software engineers for the complexities of the professional landscape. Addressing this misalignment is essential for streamlining workforce development efforts ensuring graduates are better equipped to seamlessly integrate into and contribute effectively to their roles within the software engineering domain. However, there is a current and commendable effort to address and bridge this divide, recognizing the need for a more aligned and responsive approach to Software Engineering Education [2]. So skill enhancement is pivotal in shaping successful careers, underscoring the significance of acquiring specific skills, including Soft, Life, and Technical skills. Therefore, predicting students' skill sets establishes a balance among these competencies proving essential for excelling in the dynamic and collaborative environment of software development pursuits. Additionally, recognizing the importance of skill prediction for students is essential to understand their skills, such as Soft, Life, and Technical skills.

Within the sphere of software development projects, individuals are tasked with wielding a comprehensive trio of proficiencies spanning Soft, Life, and Technical skills. Soft skills, often labeled as interpersonal or people skills, constitute the non-technical aptitudes crucial for effective interaction and navigation in social environments [3]. On a broader spectrum, life skills form an encompassing category vital for daily living, entailing personal development, self-management, and carefully through diverse life scenarios with effectiveness [4]. Technical skills, on the other hand, encompass the specific and practical abilities indispensable for crafting, testing, and maintaining software applications. These skills necessitate adept utilisation of various approaches, tools, and techniques to accomplish tasks efficiently [5]. The amalgamation of this diverse skill set transcends mere functionality; it evolves into an art and science propelling the triumphant trajectory of software development.

Machine learning, with its powerful algorithms has become instrumental in accurately predicting students' skill sets. However, the inherent black-box nature of conventional machine learning poses challenges in understanding which features contribute most to predictions and also determining the reliability of the prediction model. To address these limitations, Explainable AI (XAI) is incorporated in this study offering visualizations of crucial features through SHAP (Shapley additive explanation). The accuracy of the prediction model can be further enhanced by employing feature extraction techniques. Principal Component Analysis (PCA) serves as a feature extraction technique ensuring the extraction of high-quality and relevant features. This integrated approach enhances the accuracy and transparency of the machine learning model. The study delves into the fusion of XAI and machine learning techniques employing a PCA-based Random Forest (RF) algorithm for predicting various skill sets among computer science students. This research equips students for prosperous careers, ultimately contributing to higher success rates in software projects. By bridging the gap between the education system and the IT industry, these findings hold the potential to foster a more seamless transition for students into the professional landscape. The key points highlighted in the paper are as follows:

(1) The study introduces a predictive model incorporating a PCA-based Random Forest (RF) algorithm, leveraging Explainable AI to predict diverse skills including Soft, Life, and Technical skills.
(2) PCA serves as the feature extraction technique ensuring the extraction of high-quality and relevant features.
(3) This research uses Explainable AI in conjunction with SHAP (Shapley additive explanation) to discern critical factors influencing skillsets, offering valuable insights into the mechanisms underlying the prediction process.
(4) The study uses a customized dataset tailored to the context enhancing the relevance and specificity of the results.
(5) The proposed PCA-based Random Forest through XAI undergoes comprehensive evaluation using different metrics to gauge its effectiveness and performance under diverse conditions.

The document is structured with distinct sections. In Section 1, a brief introduction is provided on methods and strategies for predicting the abilities of students. Section 2 outlines the motivation behind the study. The associated works are summarized in Section 3. Details about the materials and methods are presented in Section 4, and Section 5 elaborates on the suggested framework. The parameter settings of the model are showcased in Section 6. The findings are discussed in Section 7. Finally, Section 8 concludes the work by presenting the investigation's results and offering suggestions for future research.

2. MOTIVATION OF STUDY

In the realm of IT industry insights, it has become evident that individuals holding degrees in software engineering often diverge significantly from the expected standards. This disparity has raised eyebrows within the academic community which traditionally strongly emphasizes software engineering subjects. The gap between the software industry and academia underscores the need for concerted efforts to bring the two close together. Recognising this imperative, it becomes crucial to identify the areas of weakness among students and proactively shape their skill sets, especially when they choose major in software engineering [6]. Acknowledging the significance of predicting students' skills is crucial for comprehending their proficiencies in Soft, Life, and Technical skills.

Numerous studies highlight the disadvantages graduates face from institutions which neglect the development of Soft, Life, and Technical skills—such as communication, teamwork, creativity, and leadership [5]. These skills are not merely advantageous but have been proven essential in the professional landscape prompting companies to invest time and resources in induction programmes for fresh graduates to cultivate these proficiencies. A noteworthy revelation from research suggests that the value of software engineering education cannot be overstated in the business world. This finding accentuates the symbiotic relationship between Software Project Management Education and the Software Engineering Industry [7, 8, 9, 10]. Consequently, undergraduate students stand to gain significantly from embracing this association, paving the way for a more seamless transition into the demands of the software engineering profession. The goal is to identify or predict the skillset that not only imparts technical knowledge but also cultivates a comprehensive set of skills essential for success in the dynamic landscape of the software engineering field.

3. RELATED LITERATURE

The precision with which machine learning models execute prediction tasks has been greatly improved by the development of Artificial Intelligence (AI) techniques. Numerous scholars have made contributions to this advancement by putting forward creative studies that focus on skill set prediction. These initiatives reflect the changing environment in which AI technologies are continually developing to meet the increasing needs of predictive analytics. The nexus between skill set prediction and machine learning has drawn the attention of academics enabling the creation of models that can more accurately identify and predict students' skills. These contributions highlight the ongoing development of AI techniques and positively impact workforce planning and personnel management strategies across various industries. The emergence of AI-driven prediction models emphasises the transformative impact of ongoing research in Artificial Intelligence offering the potential for more accurate and tailored insights into students' skill sets.

Several studies have been undertaken to discern the requisite skill set in software engineering demanded by the IT industry. Garousi V et al. [1] identify that college degree does not prepare graduates for the abilities required by the software industry. Many Software Engineering (SE) students typically encounter challenges while starting their professions. Garousi V et al. [2] observed that to address the question of how to effectively train future IT professionals, it requires highlighting the importance and filling in knowledge gaps across a number of software engineering disciplines.

Cihan P and Kalipsiz O. [4] examineS how projects created by students in software development courses are evaluated using a basic questionnaire. Its goal is to estimate both the students' hard and soft abilities. Sunindijo R.Y. [5] claims that a project's performance is significantly influenced by the variety of responsibilities that project managers have as well as their unique set of skills for assessing how well those activities will improve project efficiency in the IT industry.

Akdur D [6] presented that universities as well as businesses should make investments in skill development; companies can provide students with real-world experience while universities can adapt curricula to meet industry demands. Garcia I et al. [11] observed that the article presents the foundation of a dynamic framework which will link the requirements and goals of software development with what universities really have to offer in terms of teaching the field. This will guarantee that the software industry receives the right abilities and that students have the right possibilities to work in the sector. Mezhoudi N et al. [12] suggested that in order to fulfill the demands of the quick changing labour markets, educators might concentrate on teaching more applicable skill sets. Programme administrators may foresee and enhance their curricula to develop new capabilities for teaching, training, and reskilling both present and future employees.

Akdur D [13] presented that academics need to be aware of the competencies required to modify curricula in order to make educational programmes more successful. While experts and academics both shape and utilise SE, these two distinct realms have distinct objectives and priorities. Fang X et al. [14] observed that college graduates with an IS (Information Systems) degree might not have the ability to succeed in a beginning corporate position. The knowledge and skill sets that are now needed for a new entry-level IS hire are identified by this study.

According to Surakka S [15], the results of a brief study of instructors, learners, and IT workers in Finland have significant ramifications for computer science undergraduate programmes. Stevens D [16] observed that the report addresses certain identified constraints and future study directions expanding on prior linking work. This is achieved by greatly increasing the number of businesses from whom IT experts evaluate their abilities and involving MIS professors in the course planning process. Liebenberg J [17] presented the qualitative findings from a combination of techniques and an investigation of software engineers' opinions about the subjects they felt required to study in school. According to the report, there is a disconnect between software engineering educational institutions and business demands.

Patacsil F.F and Tablatin C.L.S [18] suggested a skills gap approach that measured the significance of the Information Technology (IT) skills barrier as viewed by learners in IT and professionals by using the participant's experiences in the internship programme. Carmen Iriarte [19] noted that the findings of this paper indicate that IT project success is predominantly influenced by soft skills. Gregory J. Skulmoski [20] explored the soft skills essential for IS project managers to thrive at different stages of the project.

Kartik N et al. [21] employed Random Forest and LSTM algorithms to improve the accuracy of forecasts of student performance. The goal of these machine learning techniques is to provide insight into their inner workings and explainability has been achieved through the use of the LIME and SHAP methodologies. Guleria P and Sood M [22] observed that the system combines the capabilities of Explainable AI (XAI) and Machine Learning (ML) to analyse educational elements that aid students in landing jobs and making the best decisions for their professional development.

Swamy V. et al. [23] put into practice five cutting-edge approaches for clarifying black-box machine learning models. They analyse and contrast their advantages when it comes to the final job of predicting student success for five massively open online courses. Nachouki M. et al. [24] suggested a framework in this study based on the random forest approach. The results demonstrated that it is possible to identify the difficult courses that students who are at risk discover.

Jayaprakash S. et al. [25] proposed the two most crucial factors: those that affect students' academic performance and help identify the pupils who are at risk. Additionally, it suggested a method known as the enhanced random forest classifier, and the goal of this method is to get greater accuracy. This study presented a comprehensive evaluation of the literature emphasizing soft skills for the achievement of IT-related projects. Petkovic D et al. [26] presented an ML framework that uses the Random Forest algorithm to analyse team activity measures and team results. The outcome demonstrated that RF can forecast SE processes and product team performance in a way that makes sense to the human eye. Md. S. and Krishnamoorthy S [27] presented context-bound cognitive skill ratings that work well for flagging student performance. An analysis of the created model reveals that this feature reduces the work required to build features for every field and may be used in a variety of courses. Lin H Y and You J [28] identified the competencies required for collaborative task. It has also been found that some characteristics have a direct impact on team projects and how well the outcomes are assessed.

Petkovic D. et al. [29] presented a novel method for forecasting and evaluating student learning about the efficiency of collaborative work in software engineering education. This is done by using the random forest (RF) technique to forecast the efficiency of student collaborative work. According to Kumar M Singh A J and Handa D. [30], the primary goal is to give readers a thorough grasp of the many data mining methods that have been applied to assessing student achievement and advancement. In reality, we aim to use the best data mining tools to raise the student's educational achievement.

Kolo D K and Adepoju S. A [31] suggested that the decision tree method be applied in this study to predict the academic progress of the pupils using a decision tree framework. It was discovered that several factors, such as the students' gender, learning motivation, and socioeconomic status, affected their performance. Makhoba L. et al. [32] propose a method for predicting students' aptitude for a scientific degree programme utilising a skill-set-based model that depicts academic performance. With a 95% accuracy rate, the random forest classifier turned out to be the most successful prediction model.

Most researchers in the field have predominantly focused on Soft skills, Technical skills, or Hard skills within software engineering, resulting in a notable gap in the exploration of L ife skills. The study recognizes the importance of Life skills in shaping well-rounded professionals and emphasizes their inclusion in predicting software engineering skill sets. However, this study also underscores an imminent gap that requires attention. Here, the relevance of prediction-related machine learning methods becomes crucial as they offer a systematic approach to anticipate and address identified skill gaps in software engineering among computer science students. Leveraging machine learning techniques provides valuable insights into evolving skill requirements, enabling proactive alignment of education and training with the dynamic demands of the software engineering field. This highlights the indispensable role of prediction-related machine learning methods in recognizing and strategically developing skills for the future.

Recognizing the scarcity of articles addressing machine learning models for predicting students' skill sets and their fundamentals, the purpose of this study is to demonstrate how important machine learning with explainable AI is for identifying and predicting skill improvement areas. Focusing particularly on Soft, Life, and Technical skills, the research endeavors to fill a notable gap in the existing works by highlighting the significance of utilising machine learning techniques in the field of education. Through illuminating the potential of these models, the study seeks to contribute valuable insights into the process of skill assessment and enhancement. Furthermore, it assists in accurately predicting whether computer science students possess Soft, Life, or Technical skills. In light of these considerations, it can be reasonably deduced that the machine learning approach using Explainable XAI emerge as the most suitable and adept methods for achieving the objectives outlined in the proposed study.

4. MATERIALS AND METHODS

4.1 DATASET DESCRIPTION

At the core of this research, the essential components of Soft, Life, and Technical skills encapsulated within the dataset. The effectiveness of the forecasting model hinges significantly on the caliber of this dataset meticulously curated from responses to a questionnaire distributed among computer science students across various colleges in Kerala, India. The dataset encompasses skill-related quiz assessments providing a comprehensive foundation for analysis. Soft skills evaluation relies on eighteen features and five quiz questions gauging resolving conflict, critical situations, inspiration, working over hours, emotional intelligence, and rearranging schedules. Life skills assessment, on the other hand, involves a questionnaire with nine features and five quiz questions focusing on demonstration, coaching, listening, interpersonal communication, and team building. For technical skills, the dataset incorporates nine features and five quiz questions covering aspects like time management, quality management, proficiency in new technologies and hands-on tools, and challenges related to cost and time management. This rich and diverse dataset is the foundation for evaluating, predicting, or classifying students' strengths and areas for improvement across the three skill sets. Tables 1 and 2 provide an insightful breakdown of the chosen criteria employed to predict Soft skills, Life skills, and Technical skills offering a structured view of the multifaceted dimensions under consideration in the domain of software engineering among computer science students.

Table 1. Features utilized to predict Soft skills

Sl No	Features to predict Soft Skills
1	Decision-Making
2	Planning
3	Teamwork Experience
4	Confident Management
5	Meeting Deadlines
6	Stress Critical Situations
7	Communicate Closed Connect
8	Boosting Creativity
9	Multitask
10	Working Over Hours
11	Opinion Differences
12	Resolving Conflict
13	Conveying Unpopular Information
14	Working Alone
15	Rearranging Schedules
16	Inspiration
17	Motivation
18	Emotional Intelligence

Table 2. Features utilized to predict Life and Technical skills

Sl No	Features to predict Life Skills	Features to predict Technical Skills
1	Demonstrator	Time Management
2	Good Leader	Quality Management
3	Good Listening	Blend Of Management Plus Technical Subject
4	Oral message communication	Gaining Extra Information
5	Team Building	New Hands-On Tools
6	Area Of Conflict Interest	Group Project Activities
7	Good Presenter	Developing Projects Alone

8	Coach Others	Cost Or Time Management Challenges
9	Interpersonal Communication	Administrative Tasks

Here are a few exemplar questions from the survey designed to evaluate students' Soft skills:

1. Do you feel assured in overseeing a group of more than 20 team members?
2. On a scale of confidence, how comfortable are you communicating within your close network or group?
3. To enhance creativity in group discussions, how confident are you in contributing lively ideas?
4. Would you opt to delegate the responsibility of decision-making to another person?
5. When faced with an unforeseen event, are you inclined to rearrange your schedule?

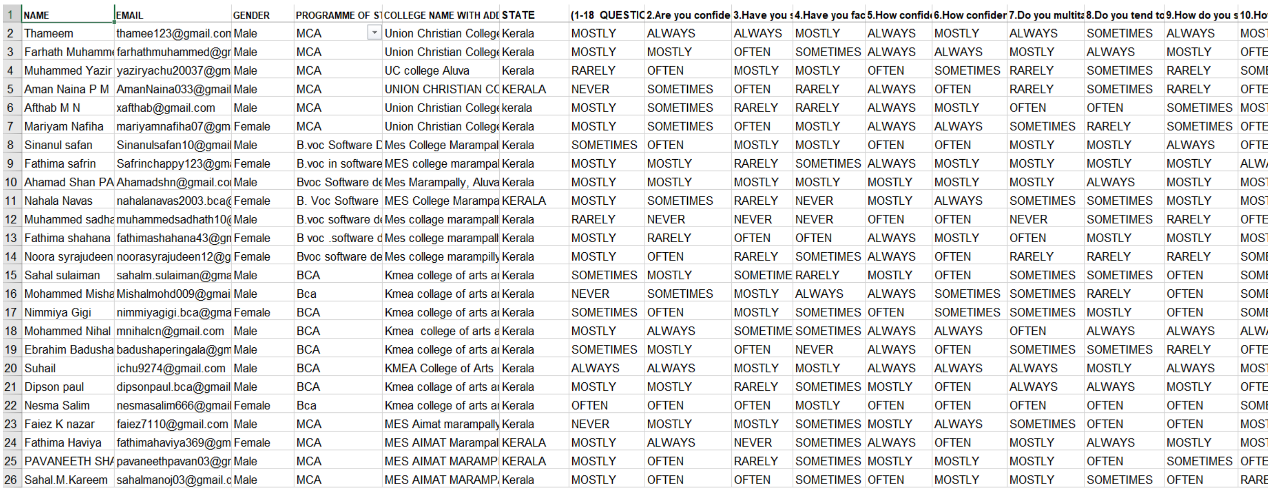
Here are a few example questions from the survey utilized to assess students' Life skills:

1. In the context of group work, how adept are you at resolving conflicts within your area of interest?
2. When engaged in group communication, do you excel as an attentive listener?
3. Are you self-reliant when it comes to delivering presentations or addressing larger groups?
4. How proficient are you in interpersonal communication?
5. In terms of motivating and constructing a core team, how effective of a leader would you consider yourself to be?

Below are a few sample questions from the survey employed to evaluate Technical skills:

1. Would you opt for a combination of management and technical subjects over a course focusing solely on management or technical aspects?
2. Are you receptive to acquiring proficiency in new hands-on tools and methodologies?
3. On a scale of confidence, how self-assured are you in independently developing a project?
4. In group projects, do you believe that challenges related to cost, computation, and time management outweigh technical challenges in significance?
5. Have you demonstrated proficiency in implementing quality management in your projects?

Figure 1 visually represents a snapshot of the dataset used in this study, offering a tangible glimpse into the structured evaluation of students' skill sets in software engineering.



1	NAME	EMAIL	GENDER	PROGRAMME OF ST	COLLEGE NAME WITH ADI	STATE	(1-18 QUESTIC	2.Are you confide	3.Have you s	4.Have you fac	5.How confid	6.How confider	7.Do you multita	8.Do you tend to	9.How do you s	10.Ho
2	Thameem	thamee123@gmail.con	Male	MCA	Union Christian College	Kerala	MOSTLY	ALWAYS	ALWAYS	MOSTLY	ALWAYS	MOSTLY	ALWAYS	SOMETIMES	ALWAYS	MOST
3	Farhath Muhamme	farhathmuhammed@gr	Male	MCA	Union Christian College	Kerala	MOSTLY	MOSTLY	OFTEN	SOMETIMES	ALWAYS	ALWAYS	MOSTLY	ALWAYS	MOSTLY	OFTE
4	Muhammed Yazir	yaziryachu20037@gm	Male	MCA	UC college Aluva	Kerala	RARELY	OFTEN	MOSTLY	MOSTLY	OFTEN	SOMETIMES	RARELY	SOMETIMES	RARELY	SOMI
5	Aman Naina P M	AmanNaina033@gmail	Male	MCA	UNION CHRISTIAN CC	KERALA	NEVER	SOMETIMES	OFTEN	RARELY	ALWAYS	OFTEN	RARELY	SOMETIMES	RARELY	OFTE
6	Afthab M N	xafthab@gmail.com	Male	MCA	Union Christian College	kerala	MOSTLY	SOMETIMES	RARELY	RARELY	ALWAYS	MOSTLY	OFTEN	OFTEN	SOMETIMES	MOST
7	Mariyam Nafiha	mariyamnafiha07@gm	Female	MCA	Union Christian College	Kerala	MOSTLY	SOMETIMES	OFTEN	MOSTLY	ALWAYS	ALWAYS	SOMETIMES	RARELY	SOMETIMES	OFTE
8	Sinanul safan	Sinanulsafan10@gmail	Male	B.voc Software D	Mes College Marampal	Kerala	SOMETIMES	OFTEN	MOSTLY	MOSTLY	OFTEN	OFTEN	MOSTLY	MOSTLY	ALWAYS	OFTE
9	Fathima safrin	Safrinchappy123@gm	Female	B.voc in software	MES college marampal	Kerala	MOSTLY	MOSTLY	RARELY	SOMETIMES	ALWAYS	MOSTLY	MOSTLY	MOSTLY	MOSTLY	ALW/
10	Ahamad Shan PA	Ahamadshn@gmail.co	Male	Bvoc Software de	Mes Marampally, Aluva	Kerala	MOSTLY	MOSTLY	MOSTLY	MOSTLY	MOSTLY	MOSTLY	MOSTLY	ALWAYS	MOSTLY	MOST
11	Nahala Navas	nahalanavas2003.bca(	Female	B. Voc Software	MES College Marampa	KERALA	MOSTLY	SOMETIMES	RARELY	NEVER	MOSTLY	ALWAYS	SOMETIMES	SOMETIMES	MOSTLY	MOST
12	Muhammed sadha	muhammedsadhath10(	Male	B.voc software de	Mes collage marampall	Kerala	RARELY	NEVER	NEVER	NEVER	OFTEN	OFTEN	NEVER	SOMETIMES	RARELY	OFTE
13	Fathima shahana	fathimashahana43@gn	Female	B voc .software d	Mes college marampall	Kerala	MOSTLY	RARELY	OFTEN	OFTEN	ALWAYS	MOSTLY	OFTEN	MOSTLY	MOSTLY	MOST
14	Noora syrajudeen	noorasyrajudeen12@g	Female	Bvoc software de	Mes college marampilly	Kerala	MOSTLY	OFTEN	RARELY	SOMETIMES	ALWAYS	OFTEN	RARELY	RARELY	RARELY	SOMI
15	Sahal sulaiman	sahalm.sulaiman@gma	Male	BCA	Kmea college of arts ai	Kerala	SOMETIMES	MOSTLY	SOMETIME	RARELY	MOSTLY	OFTEN	SOMETIMES	SOMETIMES	OFTEN	SOMI
16	Mohammed Misha	Mishalmohd009@gmai	Male	Bca	Kmea collage of arts ai	Kerala	NEVER	SOMETIMES	MOSTLY	ALWAYS	ALWAYS	SOMETIMES	SOMETIMES	RARELY	OFTEN	SOMI
17	Nimmiya Gigi	nimmiyagigi.bca@gma	Female	BCA	Kmea college of arts ai	Kerala	SOMETIMES	OFTEN	MOSTLY	SOMETIMES	OFTEN	SOMETIMES	SOMETIMES	MOSTLY	OFTEN	SOMI
18	Mohammed Nihal	mnihalcn@gmail.com	Male	BCA	Kmea college of arts a	Kerala	MOSTLY	ALWAYS	SOMETIME	SOMETIMES	ALWAYS	ALWAYS	OFTEN	ALWAYS	ALWAYS	ALW/
19	Ebrahim Badusha	badushaperingala@gm	Male	BCA	Kmea college of arts ai	Kerala	SOMETIMES	MOSTLY	OFTEN	NEVER	ALWAYS	OFTEN	SOMETIMES	SOMETIMES	RARELY	OFTE
20	Suhail	ichu9274@gmail.com	Male	BCA	KMEA College of Arts	Kerala	ALWAYS	ALWAYS	MOSTLY	MOSTLY	ALWAYS	ALWAYS	ALWAYS	MOSTLY	ALWAYS	MOST
21	Dipson paul	dipsonpaul.bca@gmail	Male	BCA	Kmea collage of arts ai	Kerala	MOSTLY	MOSTLY	RARELY	SOMETIMES	MOSTLY	OFTEN	ALWAYS	ALWAYS	MOSTLY	OFTE
22	Nesma Salim	nesmasalim666@gmail	Female	Bca	Kmea college of arts ai	Kerala	OFTEN	OFTEN	OFTEN	MOSTLY	OFTEN	OFTEN	OFTEN	OFTEN	OFTEN	SOMI
23	Faiez K nazar	faiez7110@gmail.com	Male	MCA	MES Aimat marampally	Kerala	NEVER	MOSTLY	MOSTLY	SOMETIMES	MOSTLY	ALWAYS	SOMETIMES	OFTEN	OFTEN	MOST
24	Fathima Haviya	fathimahaviya369@gm	Female	MCA	MES AIMAT Marampal	KERALA	MOSTLY	ALWAYS	NEVER	SOMETIMES	ALWAYS	OFTEN	MOSTLY	ALWAYS	MOSTLY	MOST
25	PAVANEETH SH/	pavaneethpavan03@gr	Male	MCA	MES AIMAT MARAMP	KERALA	MOSTLY	OFTEN	RARELY	SOMETIMES	MOSTLY	MOSTLY	MOSTLY	OFTEN	SOMETIMES	OFTE
26	Sahal.M.Kareem	sahalmanoj03@gmail.c	Male	MCA	MES AIMAT MARAMP/	Kerala	MOSTLY	OFTEN	OFTEN	SOMETIMES	OFTEN	MOSTLY	MOSTLY	SOMETIMES	OFTEN	RARE

Fig 1: Snapshot of dataset

In this study, a comprehensive approach is proposed for predicting software engineering skill sets, such as Soft, Life, and Technical skills, among computer science students. The methodology integrates the Random Forest (RF) algorithm, Principal Component Analysis (PCA), and Explainable AI. This combined use of RF, PCA, and XAI establishes a robust framework for skillset prediction, contributing to a more informed and insightful analysis of software engineering capabilities in the context of computer science education.

4.2.1 Random Forest

The Random Forest method [33] is widely recognized as an effective machine-learning approach for predicting academic success as evidence by numerous studies. This methodology involves constructing multiple decision trees to yield precise and reliable outcomes. Through a voting technique, the algorithm combines the results of these trees to make optimal classifications on test datasets. The random forest employs a unique strategy of randomly selecting data samples at each step to identify the most robust solution. Its effectiveness stems from its nonlinear methodology allowing for the exploration of intricate connections among different attributes. It is an effective tool for both regression and classification modelling because of this feature. The random forest method differs from other tree-based algorithms due to its unusual decision to refrain from pruning trees. Rather, it randomly divides subsets of information at each tree node improving the overall efficiency by making the forest more diverse. This approach contributes to the algorithm's robustness and effectiveness in handling diverse datasets for accurate prediction. The steps of the Random Forest Classification algorithm are given below:

Step 1: Randomly select subsets of the training data with replacement (bootstrapping).

Step 2: For each subset, build a decision tree. In each division, only evaluate a randomly selected subset of features.

Step 3: When making predictions, let each tree vote for a class. The majority class becomes the final prediction.

Step 4: Combine the predictions from all trees to make a more robust and accurate final prediction.

Step 5: At each division, use a randomly selected subset of characteristics to minimise correlations across trees and improve generalisation.

Step 6: Use data points not included in the bootstrapped sample for each tree to estimate model performance.

4.2.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a feature extraction algorithm that determines the direction of maximal variance projecting it onto a new sub-space with fewer dimensions. The covariance matrix is computed for the mean-subtracted data matrix as defined in Equation (1). Subsequently, Equation (2) yields the Eigenvector matrix V which diagonalizes the covariance matrix C utilizing D as the diagonal matrix for Eigenvalues. The principal component of the data corresponds to the Eigen-vector associated with the highest Eigenvalue. Signifying the importance, the relevant Eigenvectors provide the principal components. The model then utilizes these extracted features for making predictions enhancing both simplicity and predictive accuracy [34].

$$\boldsymbol{C} = \frac{\mathbf{1}}{\mathbf{n-1}}\boldsymbol{B} * \boldsymbol{B} \tag{1}$$

$$\boldsymbol{V^{-1}CV} = \mathbf{D} \tag{2}$$

4.3 Explainable AI (XAI)

Explainable AI [35] serves the crucial purpose of elucidating the internal processes of a model offering users a transparent understanding of methods, procedures, and outputs that are comprehensible. The term "Explainable AI" is often referred to as "White box" due to its emphasis on elucidating the model's processes. The input to this system involves training data and users are tasked with selecting a methodology for prediction based on specific requirements or application domains. This transparency ensures that users possess knowledge about the AI model's output, fostering increased trust. Armed with this knowledge, users can enhance the accuracy of outcomes, identify potential flaws in the model, and make informed decisions to refine further and improve the AI model.

4.3.1 SHAP based Explanation

Shapely Additive explanation functions as a pivotal "feature attribution mechanism" [36] in the monarchy of interpretability for machine learning models. Operating akin to a game theory method, SHAP excels in ease of calculation and offers a more intuitive interpretation. It allows explanations to be created both locally and globally exhibiting consistency across a range of data kinds and proving to be model-independent. The versatility of SHAP positions it as a valuable tool in the arsenal of Explainable AI contributing to the enhancement of model interpretability and trustworthiness.

4.4 METHOD USED FOR COMPARISON

In this study, the recommended RF-PCA model is systematically compared with prominent algorithms such as Decision Tree, Logistic Regression, Support Vector Machine, and Random Forest. This comparison investigation aims to provide insightful information about the enhanced prediction power of the RF-PCA framework when combined with the Explainable Artificial Intelligence (XAI) methodology. It comprehensively explains its efficacy in predicting software engineering skill sets among computer science students.

4.4.1 Logistic regression

Logistic regression is designed to forecast the outcome of a dependent variable that falls into distinct classes [37]. The parameter that is dependent in this instance is categorical accepting values of True or False, Yes or No, or 0 or 1. Unlike linear regression, logistic regression delivers probabilistic values from 0 to 1. Its primary utility lies in scenarios where the outcome needs to be a binary classification.

4.4.2 Decision tree

A decision tree is a visual representation that resembles a flowchart providing a structured approach to decision-making [31]. Nodes in the tree represent work requirements, branches depict decision-making criteria, and the final nodes signify the outcome of the task. The first node in the sequence, known as the tree's root node, decides how to divide the tree into segments based on the characteristics of the data. This recursive structure continues to divide the tree into branches, facilitating a comprehensive understanding of decision paths.

4.4.3 Support Vector Machine

Support Vector Machines (SVM) are widely acknowledged for their effectiveness in supervised learning classification [31]. It is particularly useful when data cannot be separated linearly or when a robust decision boundary is required to adapt effectively to new, untested data, SVM categorizes diverse types of data. By leveraging the concept of support vectors, SVM identifies the optimal hyperplane that maximally separates different classes making it a valuable tool for complex classification tasks.

5. PROPOSED FRAMEWORK

In this study, a novel approach is employed by integrating the Random Forest algorithm with the Principal Component Analysis algorithm, coupled with Explainable Artificial Intelligence, to predict diverse skill sets within the software engineering domain among computer science students. The proposed methodology encompasses several key steps including data collection, data preprocessing, feature extraction, model selection, model training, and evaluation. By combining the robustness of Random Forests with the dimensionality reduction capabilities of PCA, this methodology aims to enhance the accuracy and interpretability of skillset predictions. Incorporating explainable AI adds transparency to the decision-making process, providing insights into how the model arrives at its predictions. The sequential layout of the proposed framework is visually represented in Figure 2 offering a clear and structured overview of the analytical framework adopted for skillset prediction in the context of software engineering among computer science students.

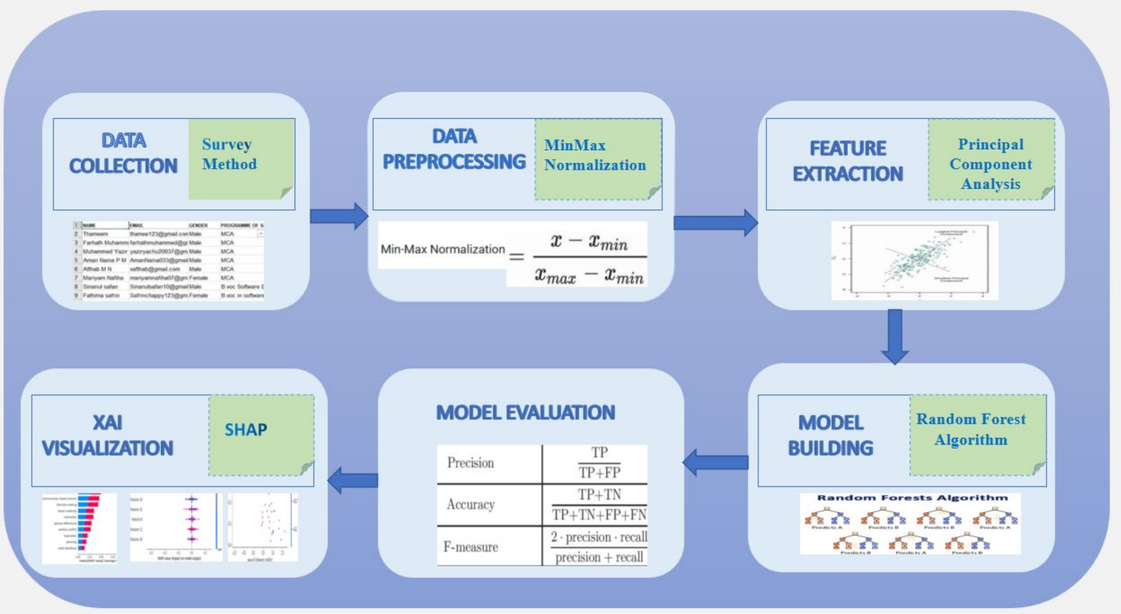

DATA COLLECTION
Survey Method
DATA PREPROCESSING
MinMax Normalization
Min-Max Normalization $= \frac{x - x_{min}}{x_{max} - x_{min}}$
FEATURE EXTRACTION
Principal Component Analysis
XAI VISUALIZATION
SHAP
MODEL EVALUATION
Precision $\frac{TP}{TP+FP}$
Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$
F-measure $\frac{2 \cdot precision \cdot recall}{precision + recall}$
MODEL BUILDING
Random Forest Algorithm
Random Forests Algorithm

Fig 2: Proposed Framework

The foundation of any prediction model is intricately tied to the quality of its data, and in this study, surveys emerged as the primary method for data collection. A meticulously crafted questionnaire served as the instrument for a survey conducted among computer science students across various colleges in India. The survey specifically targeted students who were enrolled in software engineering courses as part of their curriculum. The questionnaire designed to identify skill gaps, comprised a series of questions covering a spectrum of parameters associated with Soft skills, Life skills, and Technical skills. Each question gave respondents six options, assigning scores of 0,1,2,3,4 and 5, respectively. In order to complement the questionnaire, a quiz assessment was integrated into the survey with each skill category featuring five questions resulting in a comprehensive dataset of over a thousand samples. A score of five denoted the highest-rated skill while zero indicated the least-rated skill. The evaluation criterion for skill proficiency was set at a threshold of 60% allowing the model to accurately predict whether students possessed competence in the three categories: Soft skills, Life skills, and Technical skills. A total of one thousand and twenty-six students actively participated in completing the survey, contributing to the robustness and representativeness of the dataset. Eighty percent of the dataset is selected for training purposes while the remaining twenty percent is allocated for testing in this survey methodology.

The survey questionnaires serve as a comprehensive tool for evaluating students' skill sets in software engineering, encompassing fifty-three questions organized into three distinct parts. The first segment focuses on Soft skills, the second on Life skills, and the third on Technical skills. In assessing Soft skills, eighteen features and five quiz questions were employed to gauge each student's proficiency. Soft skills pertinent to software engineering such as managing large projects and collaborating within a software team are often underemphasized in traditional curriculum pedagogy. Life skills evaluation incorporates nine features and five quiz questions delving into areas influenced by workplace advancements including presentation skills, leadership, active listening, communication skills, and conflict resolution within specific areas of interest. Technical skills evaluation constituting the third part, involves nine features and five quiz questions that span facets like time management, quality management, adeptness with new technologies and hands-on tools, and addressing challenges related to cost and time management. This approach not only furnishes students with insights into their academic skills but also yields a comprehensive understanding of their skillset gaps. The prediction involves identifying skills in categories such as Soft, Life, and Technical skills.

5.2 Data Preprocessing

Effective data preprocessing techniques play a pivotal role in readying the dataset for the creation of machine learning models. Before constructing models, it is imperative to appropriately preprocess the collected data for training and testing enabling the models to discern underlying trends swiftly. Addressing missing values is a crucial aspect given that data with such gaps cannot effectively train a machine-learning model. Consequently, a significant portion of our time—90%—is dedicated to data preparation. In this study, one method for managing missing values entails replacing them with the average. Among the normalisation techniques investigated, min-max normalisation produced better outcomes compared to standard scaler normalisation.

MinMax normalization is a data preprocessing technique commonly used in machine learning and data analysis. This method scales and transforms numerical features within a specific range, typically between 0 and 1. The process involves determining the lowest and highest values of a variable and then normalizing each data point proportionally within that range. This normalization ensures that every feature has an equal contribution to the analysis and prevents certain variables with larger magnitudes from dominating the model training process. MinMax normalization is particularly useful when dealing with datasets containing features with varying scales, promoting a consistent and standardized representation of the data for improved model performance and interpretability. Equation (3) represents the min-max normalization equation, where min represents the smallest value in x, max is the largest value in x, and x' denotes the normalized data [38].

$$x' = \frac{x - min}{\max - min} \tag{3}$$

Feature Extraction

In certain instances, augmenting the number of features in a model may not necessarily enhance its accuracy; in fact, it can lead to a phenomenon known as overfitting. To address this challenge, feature extraction proves invaluable in optimising models and mitigating overfitting. By transforming datasets with numerous features into a modified feature domain, this technique streamlines the modelling process. In this study, Principal Component Analysis is utilised for feature extraction, aiming to identify and preserve the most significant variance directions within a high-dimensional dataset. This method is particularly beneficial in the context of datasets with numerous dimensions.

5.3 Model Training

The model introduced in this study was constructed using the Random Forest with Principal Component Analysis technique incorporating Explainable Artificial Intelligence. The implementation of the Random Forest along with PCA was achieved using the robust functionalities provided by the Python Scikit-learn (Sklearn) machine learning toolkit. This toolkit, renowned for its versatility and efficiency, played a pivotal role in seamlessly integrating the Random Forest with PCA and facilitating the utilization of explainable AI methods ensuring transparency and interpretability in how the model makes decisions.

In this investigation, the process of hyperparameter tuning utilized a randomized search approach known for its effectiveness in identifying optimal parameters. The randomised search algorithm was applied to fine-tune key parameters of the Random Forest model including the number of estimators, minimum leaf samples, maximum features, maximum depth, and the use of bootstrap. These parameters play a vital part in shaping the training process of the Random Forest algorithm. It is important to note that, especially when dealing with large datasets, the model's overall training time increases as it undergoes training and cross-validation with diverse parameter combinations [39]. The results of the randomized search showcasing the parameters with the most favorable scores are summarized in Table 3, providing insights into the configurations that contribute to the model's optimal performance.

Table 3. Optimum values for the parameters

Sl No	Name of Parameter	Value
1	Number of estimators	351

2	Minimum samples of split	5
3	Minimum samples of lea	2
4	Maximum depth	5
5	Maximum feature	'sqrt'
6	Bootstrap	False

The configuration of key hyperparameters significantly influences the behavior of the Random Forest algorithm. The number of estimators, a critical parameter dictates the quantity of individual decision trees incorporated into the random forest ensemble. Meanwhile, the minimum leaf sample parameter governs the minimum number of samples required at each decision tree's leaf node impacting the granularity of the model's predictions. In the context of decision trees, the maximum feature parameter determines the maximum number of features considered when determining optimal splits at each node, contributing to the model's flexibility and generalization. Furthermore, the maximum depth parameter influences the depth of each decision tree impacting its complexity and potential overfitting. The random forest employs bootstrapped sampling during the construction of each decision tree. This technique generates multiple datasets by randomly sampling from the original data with replacement ensuring diversity and robustness in the ensemble of trees. This method enhances the overall effectiveness of the Random Forest algorithm in capturing complex patterns and relationships within the data.

5.4 Model Evaluation

The evaluation metrics of accuracy, precision, recall, and F1-score, identified as widely employed benchmarks for classification problems, serve as crucial tools in assessing the effectiveness of the proposed framework. All together, these measures offer insightful information about how well the models are performing. The pursuit of higher values in these measurements aligns with the intention to enhance the overall functionality of the models. In an ideal scenario, a perfect classifier achieves precision and recall values both equal to one, underscoring the significance of these metrics in gauging the robustness and reliability of the suggested framework.

The concept of accuracy is defined as the ratio of accurately predicted events by the model to the total observed data, and its mathematical representation is encapsulated in Equation (4). Precision, on the other hand, quantifies the proportion of relevant observations among all retrieved observations as depicted in Equation (5). In the realm of model evaluation, recall signifies the fraction of actual positive instances that the model correctly identifies, and its mathematical representation is captured in Equation (6). For a comprehensive assessment that considers the harmonic mean of precision, recall, and the F1-score, Equation (7) offers a computational framework [25]. These equations provide a quantitative foundation for understanding and evaluating the fundamental metrics that contribute to the performance analysis of machine learning techniques.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Precision = \frac{Number\ of\ retrieved\ and\ relevant\ observations}{Total\ retrieved\ observations} = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{Number\ of\ retrieved\ and\ relevant\ observations}{Total\ relevant\ observations} = \frac{TP}{TP+FN} \quad (6)$$

$$F1\ Score = 2 * \frac{(recall * precision)}{(recall + precision)} \quad (7)$$

True Positive (TP) in a classification model corresponds to the count of accurately identified relevant observations, as defined by equations (4) to (7). Conversely, "False Positive" (FP) denotes the instances where the model incorrectly labels irrelevant observations as relevant. Conversely, "False Negative" (FN) represents relevant data mistakenly classified as irrelevant by the model. The area under the Receiver Operator Characteristic (ROC) curve known as the "Area Under the Curve" (AUC) is a pivotal metric for evaluating classification model performance. With both True Positive Rate (TPR) and False Positive Rate (FPR) ranging from 0 to 1, the AUC value ranging between 0.5 and 1 indicates model discrimination capability. Greater performance of the model is shown by a larger AUC value which reflects its ability to discriminate between positive and negative classes at different thresholds. The ROC curve which graphically depicts the disparity between true positive and false positive rates provides an accurate visual representation of the classification accuracy of the model. Essentially, the AUC measure provides a thorough evaluation of a model's efficacy with a value of 1 indicating the best classification performance.

Additionally, the study employs a confusion matrix [40] to assess the efficiency of classification models. By displaying the counts of true positive, true negative, false positive, and false negative predictions, it thoroughly assesses a machine learning algorithm's performance. These metrics show how well the model accurately categorizes occurrences in a binary classification situation. True negatives are examples of events that are accurately identified as negative, and true positives are examples of cases that are correctly classified as positive. False positives happen when cases are mistakenly classified as positive whereas false negatives happen when instances are mistakenly labelled as negative. The confusion matrix provides a thorough knowledge of the advantages and disadvantages of a model and serves as the basis for the computation of several performance measures including F1 score, accuracy, and recall. These matrices represented as two-dimensional arrays offer a comprehensive view of the model's predictions.

5.5 Explainable AI – SHAP

The study uses SHAP-based Explainable AI algorithms to uncover essential elements necessary for precise prediction. Experiments have been conducted with Explainable AI algorithms for comprehending the trained model's choice. When it comes to offering insights into the model's decision-making process, SHAP is crucial. Key characteristics are identified, and their influence on predictions is quantified providing a more detailed knowledge of the elements impacting model outputs. This all-encompassing strategy improves the predictive model's transparency and reliability making it a valuable tool for deciphering the complex correlations between input data and anticipated outcomes.

6. MODEL PARAMETER SETTINGS

The efficacy of the proposed model, Random Forest with PCA, is subjected to a comparative analysis against established benchmark model namely Logistics Regression, Decision Trees, Random Forest, and Support Vector Machine. Support Vector Machine algorithms demonstrate notable performance across both high- and low-dimensional datasets owing to the effectiveness of the kernel function. The Random Forest employed as a meta-learner leverages averaging to fit numerous decision trees to diverse dataset sub-samples, thereby enhancing prediction accuracy and mitigating overfitting risks [41]. The parameter settings for each of these machine learning models utilized for comparison are outlined in Table 4 providing a comprehensive overview of the configuration employed in the comparative evaluation process.

Table 4. The parameters used by the models used for comparison

Models	Parameters	Number or Type
SVM	Kernel function	Radial basis function (RBF
DT	Maximum depth of base estimator	10
	Minimum samples of leaf of base estimator	5
	Learning rate	0.1
	Number of estimators	300
	Maximum depth of base estimator	10
RF	Number of estimators	3
	Minimum samples of split	5
	Minimum samples of leaf	2
	Maximum feature	Sqrt
	Bootstrap	FALSE

7. RESULTS AND DISCUSSIONS

This study advocates for adopting Random Forest with Principal Component Analysis (RF-PCA) with XAI as a robust approach for predicting students' skill sets. PCA is utilized for feature extraction, enhancing the model's ability to discern relevant patterns in the data. The assessment of model performance incorporates key metrics such as accuracy, precision, recall, F1-score, and the Area Under the Curve. The recommended RF-PCA model is then systematically compared against prominent algorithms, including Decision Tree, Logistics Regression, Support Vector Machine, and Random Forest. With regard to the RF-PCA's greater predictive powers over the XAI model, this comparison analysis seeks to provide light on how well it captures and interprets students' skill sets.

7.1 Assessment of Soft, Life, and Technical skills

The outcomes of the Soft, Life, and Technical skill assessment are presented in Tables 5, 6, and 7. The comparative analysis reveals that Random Forest, Logistic Regression, Decision Tree, and Support Vector Machine exhibit similar performance levels. Notably, the Random Forest with Principal Component Analysis stands out, achieving an impressive accuracy of 86.56% for Soft skills, 84.62% for Life skills, and 87.05% for Technical skills. These metrics collectively underscore the model's exceptional generalizability and its ability to predict Soft, Life, and Technical skills with a high degree of accuracy. The findings emphasize the efficacy of the Random Forest with PCA approach as a robust model for assessing and predicting Soft, Life, and Technical skills offering valuable insights into individuals' proficiency in this domain.

Table 5. Comparison of Accuracy, Precision, Recall, F1-score, and AUC values of different models for predicting Soft skills

Model Employed	Accuracy	Precision	Recall	F1-Score	AUC value
LogR	0.7014	0.7166	0.7678	0.7413	0.6929
DT	0.8059	0.7920	0.8839	0.8354	0.7958
SVM	0.8009	0.7812	0.8928	0.8333	0.7958
RF	0.8507	0.8202	0.9375	0.8750	0.8395
RF-PCA	**0.8656**	**0.8244**	**0.9642**	**0.8888**	**0.8529**

Table 6. Comparison of Accuracy, Precision, Recall, F1-score, and AUC values of different models for predicting Life skills

Model Employed	Accuracy	Precision	Recall	F1-Score	AUC value
LogR	0.6069	0.5700	0.6129	0.5906	0.6073
DT	0.8109	0.8125	0.7959	0.8041	0.8105
SVM	0.6716	0.6310	0.6989	0.6632	0.8243
RF	0.8208	0.7786	0.9134	0.8407	0.8175
RF-PCA	**0.8462**	**0.8252**	**0.8673**	**0.8457**	**0.8462**

Table 7. Comparison of Accuracy, Precision, Recall, F1-score, and AUC values of different models for predicting Technical skills

Model Employe	Accuracy	Precision	Recall	F1-Score	AUC value
LogR	0.7064	0.7289	0.7222	0.7255	0.7051
DT	0.8308	0.7747	0.9052	0.8349	0.8347
SVM	0.8258	0.7542	0.9368	0.8356	0.8347
RF	0.8606	0.8474	0.9259	0.8849	0.8661
RF-PCA	**0.8705**	**0.8442**	**0.9537**	**0.8956**	**0.8747**

The effectiveness of the RF-PCA model is comprehensively evaluated using a diverse set of statistical metrics providing a thorough understanding of its overall performance. This study employs various evaluation measure including accuracy, precision, recall, F1-score, and AUC, to assess the RF-PCA model's predictive capabilities. Figure 4 visually illustrates the comparative performance of the proposed model using accuracy against others in predicting Soft, Life, and Technical skills, showcasing its superior capabilities. The precise visual representations in these figures unequivocally demonstrate the RF-PCA model's outperformance over alternative models, establishing it as a robust and effective choice for predicting diverse skill sets.

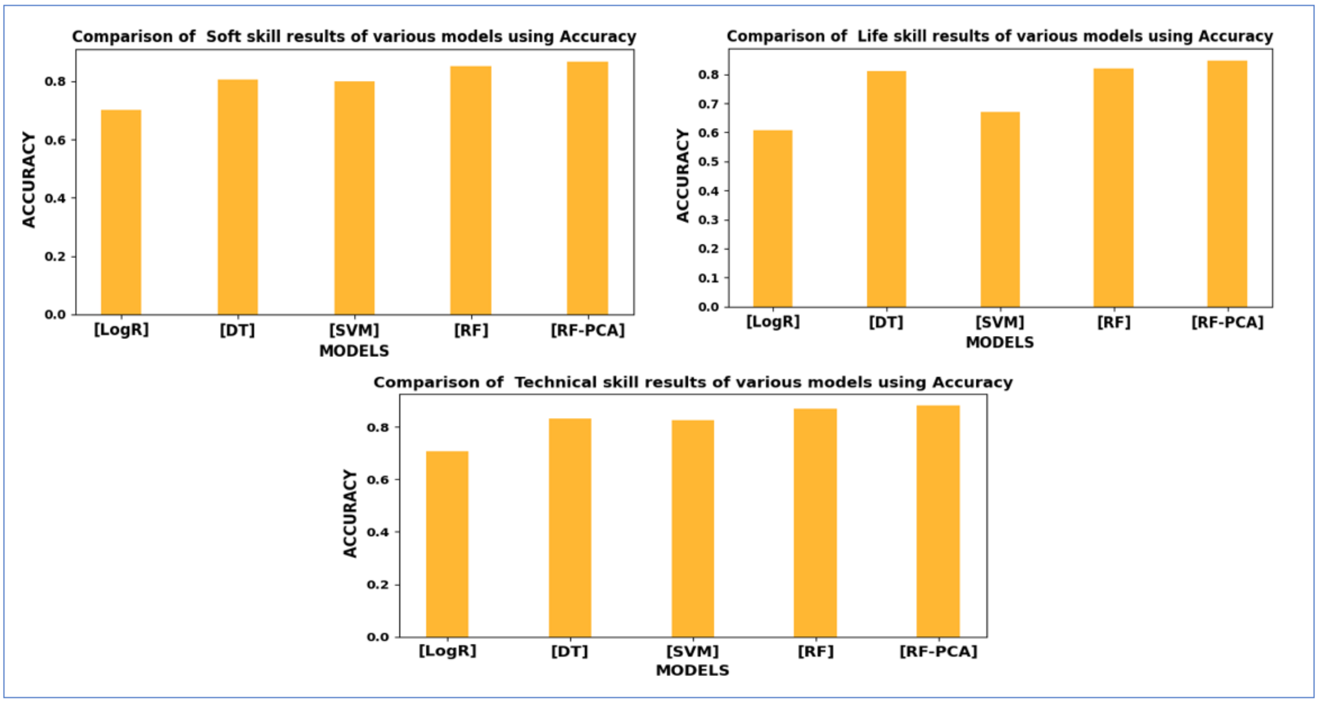
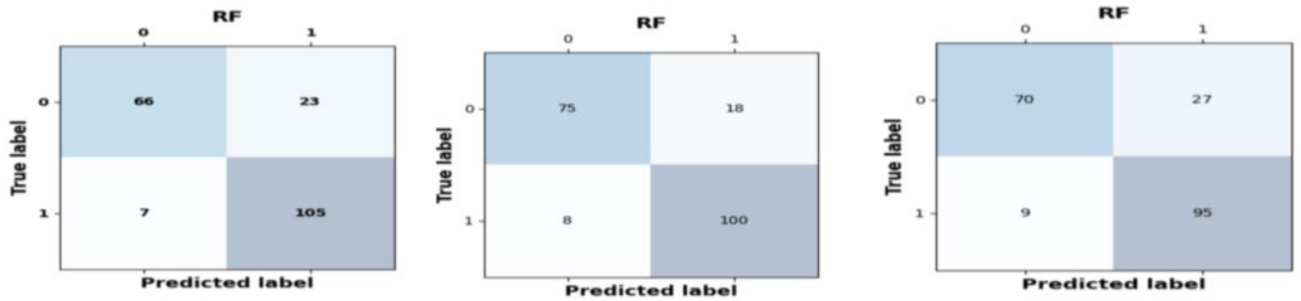


Fig 4: Accuracy values for prediction of Soft, Life, and Technical skills

7.2 Confusion Matrix of Suggested Model

The confusion matrix serves as a valuable tool for assessing the performance of classifiers particularly when employing neural networks on training datasets. It provides insights into the true and false value predictions offering a comprehensive assessment of model accuracy. The versatility demonstrated in achieving high accuracy across diverse skill domains underscores the effectiveness of RF-PCA with XAI as a robust model for skill set prediction. In Figure 5, the confusion matrix visually represents predictions for Soft, Life, and Technical skills using the RF-PCA classifier. Out of 201 predictions, the classifier accurately predicted 171 instances for Soft skills, while 30 predictions were incorrect. For Life skills there were 175 true predictions and 26 incorrect predictions. Regarding Technical skills, the classifier made 165 true predictions with 36 incorrect predictions. This breakdown provides a detailed assessment of the classifier's performance across different skill categories.



Confusion matrix for Soft skills

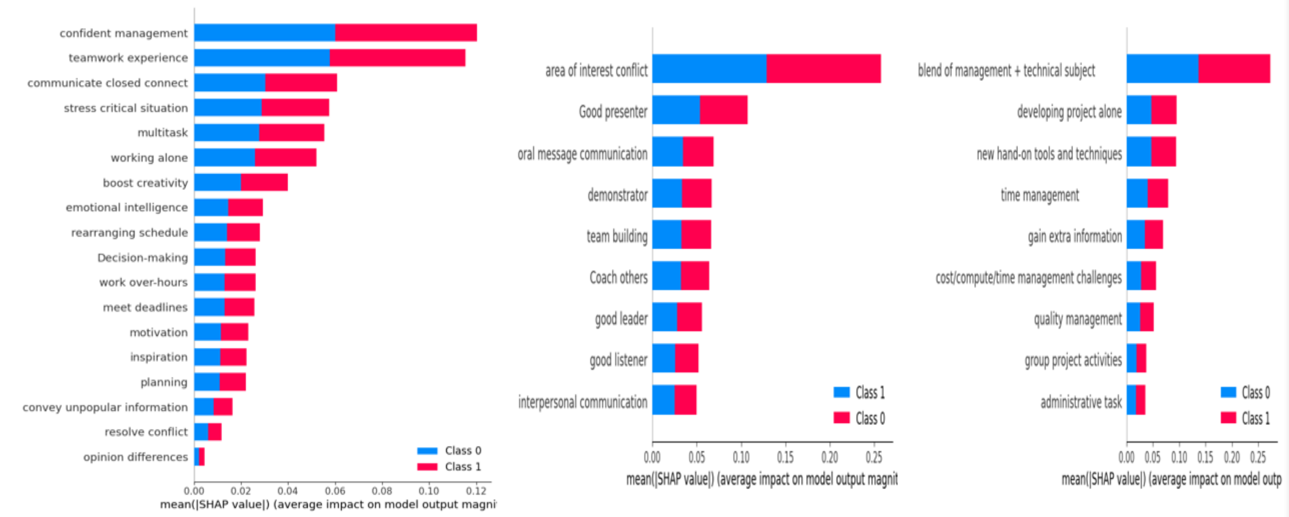
Confusion matrix for Life skills

Confusion matrix for Technical skills

Fig 5: Confusion matrix for Soft, Life, and Technical skills

7.3 Explainability of the proposed model using SHAP

SHAP enhances the interpretability of individual predictions by calculating the significance values associated with each feature [42]. The combined degree of feature importance embodied in SHAP values adhere to three crucial properties: "accuracy, consistency, and missingness."



Feature explainability for Soft skills

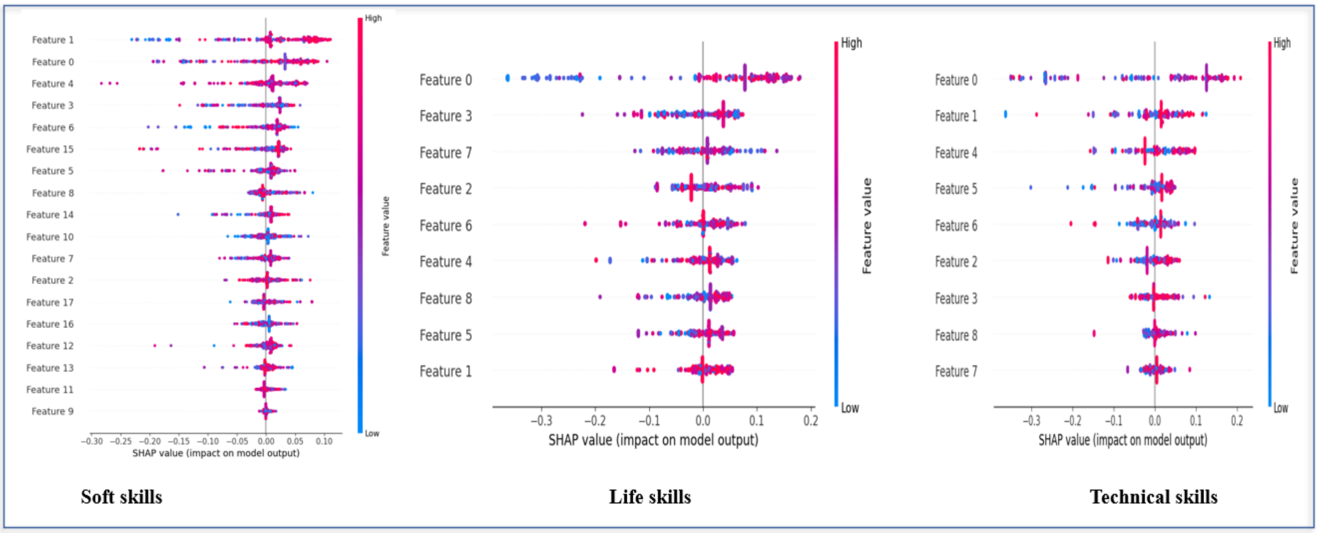
Feature explainability for Life skills

Feature explainability for Technical skills

Fig 6: Visualization of SHAP feature importance measured as mean absolute Shapley values for Soft, Life, and Technical skills

This study showcases a comprehensive approach, encompassing the training of a Random Forest model, the implementation of Principal Component Analysis for effective dimensionality reduction, and the utilization of SHAP to provide interpretability to the model's predictions. The analysis includes the creation of a summary plot and the optional generation of a force plot for a specific prediction. In Figure 6, we delve into extracting the most significant features from the RF-PCA model explicitly focusing on Soft, Life, and Technical skills.

In Soft skills, the study reveals that attributes such as confident management and teamwork experience emerge as pivotal features exerting a substantial impact on the model's predictions. In the Life skill domain, the investigation points towards the prominence of the feature related to areas of interest conflict identifying it as a key determinant. Transitioning to the Technical skill category, the analysis underscores the significance of the feature denoting a blend of management and technical subjects, establishing it as the most crucial factor influencing the model's predictive outcomes. This nuanced exploration provides valuable insights into the specific features driving the model's decision-making process across different skill categories.



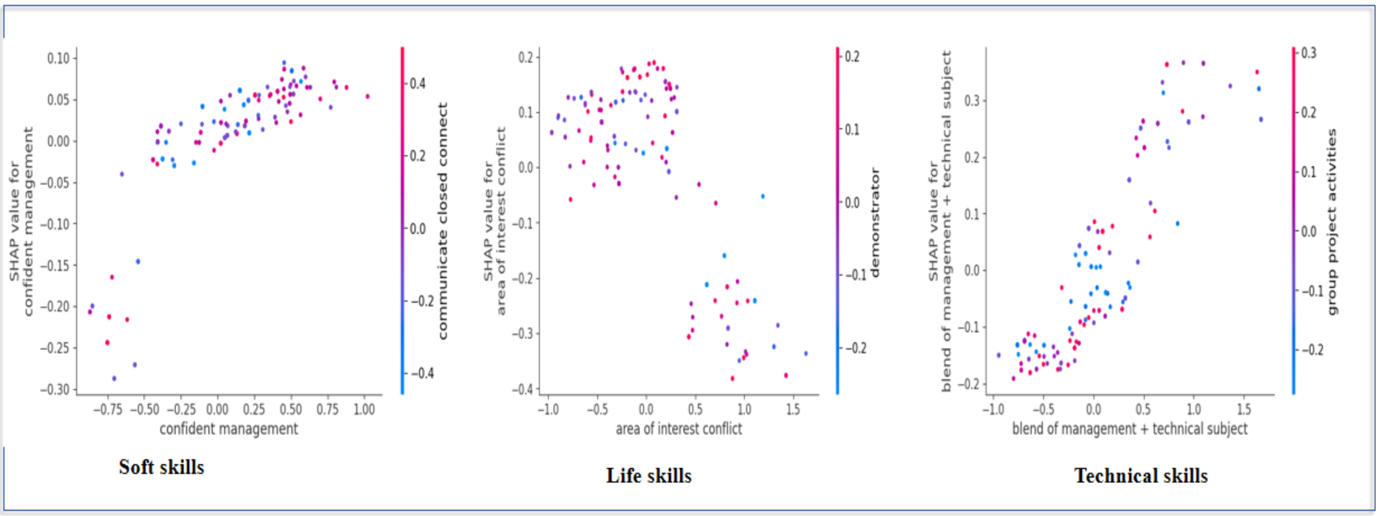
Soft skills

Life skills

Technical skills

Fig 7: SHAP summary plot for Soft, Life and Technical skills

A summary plot [43] specifically for the positive class's SHAP values and the corresponding feature values in the test set. The plot will show the distribution and impact of each feature on the model's predictions for instances in the test set specifically focusing on the positive class. As illustrated in Figure 7, the color scheme delineates the presence and absence of features, with blue signifying absence and red indicating presence. The analysis reveals that the presence of feature 0 elicits a positive contribution to the prediction of Soft, Life, and Technical skills. Conversely, the absence of this feature is associated with a negative impact on the prediction outcomes. The amalgamation of the Random Forest model with PCA coupled with the interpretative power of the SHAP algorithm, enriches the decision-making process for skill prediction among students in the domain of software engineering.



Soft skills

Life skills

Technical skills

Fig 8: SHAP dependence plot for the specific features of Soft, Life, and Technical skills

The dependency plot [44] shown in Figure 8, offers important information about how the feature affects the outcome of the model for the specified class. In exploring Soft skills, a targeted analysis is conducted by creating a dependence plot for the specific feature 'confident management.' This plot serves as a visual tool to comprehend the relationship between the chosen feature and the model's output. By utilizing SHAP values, the dependence plot sheds light on how variations in the feature 'confident management' influence predictions within the Soft skill category. The adaptability of the analysis is emphasized allowing adjustments to the feature of interest as needed to refine the investigation.

Shifting the focus to Life skills, a distinct feature, 'area of interest conflict,' is singled out for a similar exploration. A dependence plot is generated to visually interpret how fluctuations in this specific feature contribute to the SHAP values for the positive class across instances in the test set. The resulting plot offers a dynamic representation of the interplay between the chosen feature and model predictions within the Life skill domain. This targeted analysis enhances the interpretability of the model's behavior in relation to the selected feature.

For the Technical skill category, the analysis centers around the chosen feature 'blend of management + technical subject.' A dependence plot is crafted to illuminate the relationship between this feature and its corresponding SHAP values. The plot effectively illustrates how the selected feature's value affects the model's output providing nuanced insights into the feature's impact on predictions within the realm of Technical skills. This approach offers a comprehensive understanding of the model's decision-making process, contributing valuable information for informed analysis and decision support.

8. CONCLUSION

In this study, the combination of Random Forest and Principal Component Analysis utilizing explainable Artificial Intelligence is used to construct a model for predicting students' skill sets encompassing Soft, Life, and Technical skills. The performance of this proposed model is systematically compared to other well-established machine learning methods including Support Vector Machines, Decision Trees, and Logistic Regression. The experimental findings unequivocally demonstrate the superiority of the Random Forest with PCA algorithm using XAI outperforming baseline models in predicting software engineering skill sets among computer science students. Its significant capacity for generalization underscores its potential to address skill gaps and guide skill-based education initiatives in software engineering programs. The objective is to get students ready for the job market enhance the success rate of software development endeavors, and bridge the knowledge gap between educational institutions and the software sector.

This study adds value to the educational landscape by pinpointing student skill gaps. It achieves this by predicting individual skill sets including Soft, Life, and Technical skills, enabling the identification of specific areas where each student may need improvement. Consequently, these findings pave the way for targeted and tailored skill-based education initiatives aimed at addressing the unique needs of each student. Identifying skill-rich students enhances the teaching-learning process within institutions fostering advancements in the educational system. Integrating skill set assessments and predictions benefits both teachers and students promoting a culture of critical self-assessment and autonomy in the learning journey. This methodology, offering a more nuanced and accurate evaluation than traditional methods, enables better-informed decisions for educational institutions and employers in selecting and cultivating skilled software engineers. Moreover, it provides educators with valuable insights into the personalities and assets of their students.

To address the situation where a student is predicted to lack soft, life, or technical skills according to the Random Forest with PCA model, it is crucial to implement a targeted intervention approach. This involves the creation of personalized skill development plans that specifically target identified areas of weakness. Tailored resources and training programs should be provided to address these weaknesses. Additionally, a system of continuous monitoring and feedback mechanisms needs to be established to track the progress of skill enhancement efforts. The point of impact for acquiring soft, life, or technical skills can be found in the implementation of a comprehensive educational strategy. This strategy should include the integration of practical, hands-on experiences, collaborative projects, and real-world applications into the curriculum. Fostering a supportive learning environment is equally essential, encouraging extracurricular activities and providing mentorship opportunities. By combining these elements, a well-rounded skill set can be nurtured among computer science students, promoting their overall growth and preparedness for the demands of the field. This paradigm shift towards deep learning-based feature extraction and prediction models heralds a future where skill assessments are not only more precise but also more aligned with the dynamic and multifaceted nature of evolving skill landscapes.

REFERENCES

[1] Garousi, V., Giray, G., Tuzun, E., Catal, C. and Felderer, M., 2019. Closing the gap between software engineering education and industrial needs. *IEEE software*, *37*(2), pp.68-77.Belzer K 2001 Project management: Still more art than science, *In PM Forum Featured Papers* (pp. 1-6).

[2] Garousi, V., Giray, G., Tüzün, E., Catal, C. and Felderer, M., 2019. Aligning software engineering education with industrial needs: A meta-analysis. *Journal of Systems and Software*, *156*, pp.65-83.

[3] Belzer K 2001 Project management: Still more art than science, *In PM Forum Featured Papers* (pp. 1-6).

[4] Cihan P and Kalipsiz O 2014 Evaluation of students' skills in software project. *TEM Journal*, *3*(1), p.42.

[5] Sunindijo R Y 2015 Project manager skills for improving project performance. *International Journal of Business Performance Management*, *16*(1), pp.67-83.

[6] Akdur, D., 2022. Analysis of software engineering skills gap in the industry. *ACM Transactions on Computing Education*, *23*(1), pp.1-28.

[7] Begel A and Simon B 2008 Struggles of new college graduates in their first software development job. In *Proceedings of the 39th SIGCSE technical symposium on Computer science education* (pp. 226-230).

[8] Begel A and Simon B 2008 Novice software developers, all over again. In *Proceedings of the fourth international workshop on computing education research* (pp. 3-14).

[9] Gnatz M, Kof L, Prilmeier F and Seifert T 2003 A practical approach of teaching software engineering. In *Proceedings 16th Conference on Software Engineering Education and Training, (CSEE&T 2003).* (pp. 120-128). IEEE.

[10] Yeh R T 2002 Educating future software engineers. *IEEE Transactions on education*, *45*(1), pp.2-3.

[11] Garcia I, Pacheco C and Coronel N 2010 Learn from practice: defining an alternative model for software engineering education in Mexican universities for reducing the breach between industry and academia. In *Proceedings of the International Conference on Applied Computer Science* (pp. 120-124).

[12] Mezhoudi, N., Alghamdi, R., Aljunaid, R., Krichna, G. and Düştegör, D., 2023. Employability prediction: a survey of current approaches, research challenges and applications. *Journal of Ambient Intelligence and Humanized Computing*, *14*(3), pp.1489-1505.

[13] Akdur, D., 2019, June. The design of a survey on bridging the gap between software industry expectations and academia. In *2019 8th Mediterranean Conference on Embedded Computing (MECO)* (pp. 1-5). IEEE.

[14] Fang, X., Lee, S. and Koh, S., 2005. Transition of knowledge/skills requirement for entry-level IS professionals: An exploratory study based on recruiters' perception. *Journal of Computer Information Systems*, *46*(1), pp.58-70.

[15] Surakka, S., 2007. What subjects and skills are important for software developers?. *Communications of the ACM*, *50*(1), pp.73-78.

[16] Stevens, D., Totaro, M. and Zhu, Z., 2011. Assessing IT critical skills and revising the MIS curriculum. *Journal of Computer Information Systems*, *51*(3), pp.85-95.

[17] Liebenberg, J., Huisman, M. and Mentz, E., 2014. Knowledge and skills requirements for software developer students. *International Journal of Computer and Information Engineering*, *8*(8), pp.2612-2617.

[18] Patacsil, F.F. and Tablatin, C.L.S., 2017. Exploring the importance of soft and hard skills as perceived by IT internship students and industry: A gap analysis. *Journal of Technology and Science education*, *7*(3), pp.347-368.

[19] Iriarte C and Bayona Orè S 2018 Soft skills for it project success: A systematic literature review.In *Trends and Applications in Software Engineering: Proceedings of the 6th InternationalConference on Software Process Improvement (CIMPS 2017) 6* (pp. 147-158). Springer International Publishing.

[20] Skulmoski G J and Hartman F T 2010 Information systems project manager soft competencies:A project-phase investigation. *Project Management Journal*, *41*(1), pp.61-80.

[21] Kartik, N., Mahalakshmi, R. and Venkatesh, K.A., 2023. XAI-Based Student Performance Prediction: Peeling Back the Layers of LSTM and Random Forest's Black Boxes. *SN Computer Science*, *4*(5), p.699.

[22] Guleria, P. and Sood, M., 2023. Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling. *Education and Information Technologies*, *28*(1), pp.1081-1116.

[23] Swamy, V., Radmehr, B., Krco, N., Marras, M. and Käser, T., 2022. Evaluating the explainers: black-box explainable machine learning for student success prediction in MOOCs. *arXiv preprint arXiv:2207.00551*.

[24] Nachouki, M., Mohamed, E.A., Mehdi, R. and Abou Naaj, M., 2023. Student Course Grade Prediction Using the Random Forest Algorithm: Analysis of Predictors' Importance. *Trends in Neuroscience and Education*, p.100214.

[25] Jayaprakash, S., Krishnan, S. and Jaiganesh, V., 2020, March. Predicting students academic performance using an improved random forest classifier. In *2020 international conference on emerging smart computing and informatics (ESCI)* (pp. 238-243). IEEE.

[26] Petkovic D, Sosnick-Pérez M, Huang S, Todtenhoefer R, Okada K, Arora S, *et al.* 2014 Setap:Software engineering teamwork assessment and prediction using machine learning. In *2014 IEEEfrontiers in education conference (FIE) proceedings* (pp. 1-8). IEEE.

[27] Md S and Krishnamoorthy S 2022 Student performance prediction, risk analysis, and feedbackbased on context-bound cognitive skill scores. *Education and Information Technologies*, *27*(3),pp.3981-4005.

[28] Lin H Y and You J 2021 Teamwork-performance prediction by using soft skills and technological savvy skills. *Journal of University Teaching & Learning Practice*, *18*(8), p.09.

[29] Petkovic D, Okada K, Sosnick M, Iyer A, Zhu S, Todtenhoefer R, *et al.* 2012 Work in progress: a machine learning approach for assessment and prediction of teamwork effectiveness in software engineering education. In *2012 frontiers in education conference proceedings* (pp. 1-3). IEEE.

[30] Kumar M, Singh A J and Handa D 2017 Literature survey on student's performance prediction in education using data mining techniques. *International Journal of Education and Management Engineering*, *7*(6), pp.40-49.

[31] Kolo D K and Adepoju S A 2015 A decision tree approach for predicting students academic performance.

[32] Makhoba L, Jadhav A, Sixhaxa K and Ajoodha R 2022 Evaluation of Student Skill-Sets as Predictors of Success at Higher Education Institutions. In *Proceedings of International Conference on Communication and Computational Technologies: ICCCT 2022* (pp. 585-600). Singapore: Springer Nature Singapore.

[33] Cutler A, Cutler D R and Stevens J R 2012 Random forests *Ensemble machine learning:Methods and applications*, pp.157-175.

[34] Gupta I, Sharma V, Kaur S and Singh A K 2022 PCA-RF: an efficient Parkinson's disease prediction model based on random forest classification. *arXiv preprint arXiv:2203.11287*.

[35] Saranya, A. and Subhashini, R., 2023. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decision analytics journal*, p.100230.

[36] Van den Broeck, G., Lykov, A., Schleich, M. and Suciu, D., 2022. On the tractability of SHAP explanations. *Journal of Artificial Intelligence Research*, *74*, pp.851-886.

[37] Boateng E Y and Abaye D A 2019 A review of the logistic regression model with emphasis on medical research. *Journal of data analysis and information processing*, *7*(4), pp.190-207.

[38] Jaseena K U and Kovoor B C 2021 A Wavelet-based hybrid multi-step Wind Speed Forecasting model using LSTM and SVR. *Wind Engineering*, *45*(5), pp.1123-1144.

[39] Wali, S. and Khan, I., 2023. Explainable AI and random forest based reliable intrusion detection system. *Authorea Preprints*.

[40] Makhoba L, Jadhav A, Sixhaxa K and Ajoodha R 2022 Evaluation of Student Skill-Sets as Predictors of Success at Higher Education Institutions. In *Proceedings of International Conference on Communication and Computational Technologies: ICCCT 2022* (pp. 585-600). Singapore: Springer Nature Singapore.

[41] Nair A 2019 Parameter tuning with grid search: A hands-on introduction. *Analytics India Magazine*

[42] Siemers, F.M. and Bajorath, J., 2023. Differences in learning characteristics between support vector machine and random forest models for compound classification revealed by Shapley value analysis. *Scientific Reports*, *13*(1), p.5983.

[43] DataCamp.,2023. Explainable AI: Understanding and trusting machine learning models.

[44] shap., 2018. Census income classification with scikit-learn.