IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 15, 2020, accepted February 19, 2021, date of publication February 23, 2021, date of current version March 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3061368

Explainable Student Performance Prediction Models: A Systematic Review

RAHAF ALAMRI[1] AND BASMA ALHARBI[2]
[1]Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Jeddah 21493, Saudi Arabia
[2]Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah 21493, Saudi Arabia

Corresponding author: Basma Alharbi (bmalharbi@uj.edu.sa)

This work was supported by the University of Jeddah, Saudi Arabia, under Grant UJ-02-079-DR.

ABSTRACT Successful prediction of student performance has significant impact to many stakeholders, including students, teachers and educational institutes. In this domain, it is equally important to have *accurate* and *explainable* predictions, where accuracy refers to the correctness of the predicted value, and explainability refers to the understandability of the prediction made. In this systematic review, we investigate explainable models of student performance prediction from 2015 to 2020. We analyze and synthesize primary studies, and group them based on nine dimensions. Our analysis revealed the need for more studies on explainable student performance prediction models, where both accuracy and explainability are properly quantified and evaluated.

INDEX TERMS Explainable artificial intelligence, explainable machine learning, student performance models, systematic literature review.

I. INTRODUCTION

Predicting students' performance in a certain course [1], [2] or an entire program [3] is an important task to many stakeholders; including students themselves, teachers, and academic institutes. Applications of student performance prediction has proven to be useful to predict at-risk students [4] and dropout rates [5]. Additionally, it is used to build early warning systems [6] and customized recommendation systems [7] to improve the students' learning experience.

Figure 1, inspired by [8], visualizes the potential stakeholders who will benefit from a student performance prediction model. The figure depicts four main stakeholders; students (i.e., the affected users by the model), academic advisors (i.e., the end users of the model), regularity bodies, and AI/ML system builders. For these stakeholders, a model with high prediction accuracy is key to its success. For example, a model that can predict whether a student will pass or fail a certain course with a 90% accuracy can be used within a system that recommends courses to students [9]. However, high accuracy is not the only factor critical to the success of such models. In this setting, it is important to trust the model, where trust here means that the stakeholders can understand the reasons behind system predictions. Thus, in the education domain, it is equally important to have accurate predictions as well as explainable ones.

The interest on eXplainable Artificial Intelligence (XAI) and explainable machine learning goes back to the 60's [10]. The advances in both machine learning and deep learning shifted the focus from explainable models (i.e., white-box models) to more deep and black-box models, which solved many challenging problems and unleashed the potential to many more interesting applications [11]. However, the recent release of the General Data Protection Regulation (GDPR) re-emphasized the importance of explainable and trustworthy AI [12]. The regulation gives individuals the right to obtain explanations of predictions made by a model. This inherently means that black-box models and deep learning ones cannot be utilized in areas where the decision affect individuals, unless these decisions can be explained!

Going back to the education sector, explainable models are of significant value to many stakeholders. Consider the task of student performance prediction as an example. In this task, at least four different stakeholders will benefit from an explainable model, and the benefits to each group varies. Decision made by a prediction algorithm in this case will affect students directly, especially if a recommendation system is built on top of such predictors. Students in this case

The associate editor coordinating the review of this manuscript and approving it for publication was Venkateshkumar M.

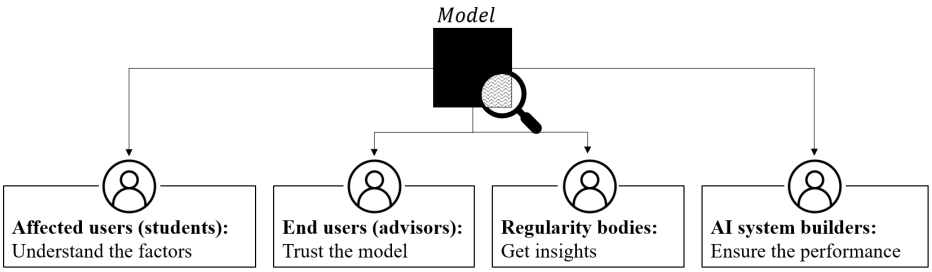


FIGURE 1. Four potential stakeholders (users) of an explainable model in the student performance prediction domain. The potential stakeholders are: 1) students who represent the *affected* users by the model, 2) advisors who are the primary end users of the model, 3) regularity bodies which may include other users in the educational institute such as the scientific departments or college heads, and 4) AI system builders who train and evaluate the prediction model.

will benefit greatly if the model explained the factors behind the decision made. For example, if it was predicted that student x will fail a certain course, an explainable model will justify this prediction and explain the factors behind it, in a language understandable by the student. Advisors are another group of beneficiaries, where they can use the predictions of a system to recommend courses to students based on their strengths. In this case, justifying the reasons behind these predictions will enable advisors to trust the model, and thus follow its predictions. Additionally, insights obtained from such explainable models can help regularity bodies and management at various levels to make improvements to current program plans. Lastly, AI system builders would be able to ensure the quality of the trained model, if they can actually 'look' inside it and verify the correctness of the inferred patterns.

Given the importance of explainable student performance prediction models, the objective of this systematic review is to study and synthesize recent work in this domain. Specifically, we review articles published between 2015 and 2020, that utilizes explainable machine learning models in student performance prediction. The main contributions of this work are summarized below:

- Draw attention to the importance of developing/employing explainable models to educational data mining in general, and student performance prediction in specific.
- Investigate and synthesize recent studies on explainable student performance prediction models.
- Identify state-of-the-art in explainable student performance prediction models.
- Highlight existing limitations and potential directions for future work.

The rest of this paper is organized as follows. Section II describes the literature review, which includes a background section and a related work section. Section III describes the research goal, questions, and methodology. Sections IV and V discuss the results and the limitations of the work. Finally, Section VI provides concluding remarks.

II. LITERATURE REVIEW

A. BACKGROUND

This section provides background and defines the key terminologies that will be used throughout the systematic review. To be consistent with existing work, we synthesized recent systematic reviews that cover two main topics: student performance prediction models, and eXplainable Artificial Intelligence. The next subsections define the main terminologies in each topic respectively.

1) STUDENT PERFORMANCE MODELS

To be consistent with existing work, we utilized terminologies and taxonomy similar to [13], [14]. Five dimensions are used to categories student performance models. The first three dimensions describe the general context of the problem, which are: education level, performance level, and problem type. The next two dimensions describe the input to the model, which are: predictors and predictors type. Detailed description of each dimension is provided next.

- **Educational Level:** This dimension categories research on student performance prediction based on the context of the data, which may include 1) specific performance level data, such as the course or program name, and 2) student level, which may range from K-12 up to graduate students.
- **Performance Level:** This dimension categories research on student performance prediction based on the level of predicted performance. In general, predicted performance can be at an assessment level, at a course level or at a program level. For example, a study that aims at predicting a student's grade at the end of the course, is labeled as course-level performance. Similarly, a study that aims at predicting students who will drop out of the college is labeled as a program-level performance.
- **Problem Type:** This dimension denotes the type of the prediction problem, which can either be classification or regression. Classification problems are further divided into two subcategories; binary and multi-way classifications.

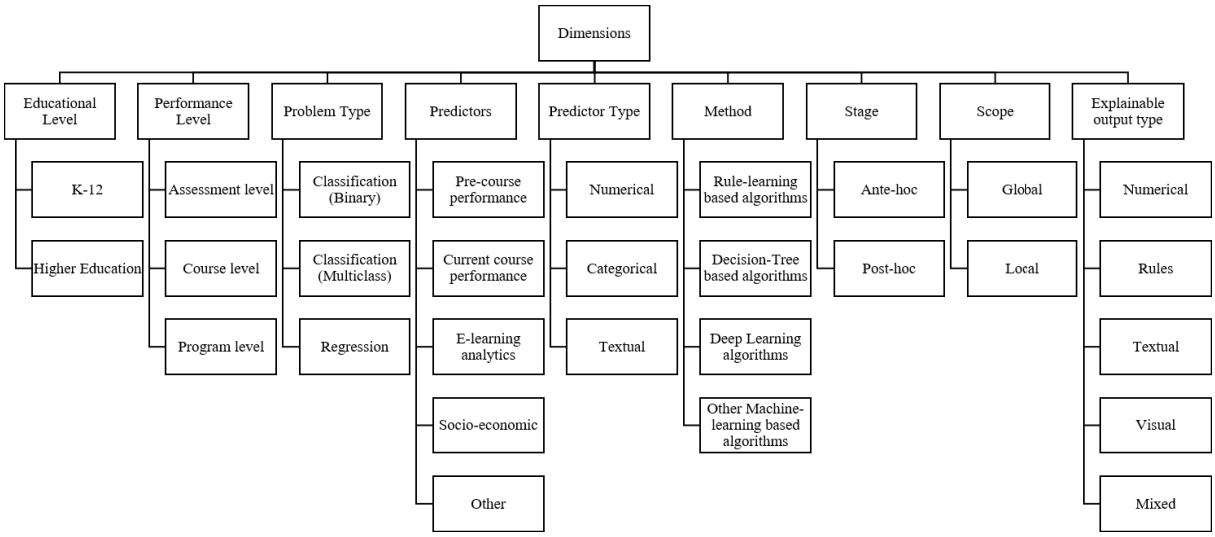


FIGURE 2. Taxonomy of the different dimensions used to categorized explainable models of student performance.

- **Predictors:** This dimension includes broad grouping of student performance predictors (input) which are: course data and student data. Course data can be further divided into pre-course data and current course data. Student data can be further divided into demographic data, personality metrics and engagement levels.
- **Predictors' Type:** This dimension refers to the type of input used which can be numerical, categorical, textual or time series.

2) EXPLAINABLE MODELS

When it comes to XAI, many terms are used interchangeably such as transparency, interpretability, and understandability to denote explainability [15]. In fact, synthesizing available literature reveals that there is no one common and clear definition of explainability in the machine learning context [15], [16]. Machine learning models are often categorized as either white-box or black-box, to indicate the transparency of the learned models [17]. In this context, black-box models refer to models that are either very complex functions or proprietary functions [17]. Deep learning models are classic examples of black-box models. White-box models, on the other hand, refer to models that can be understood by a human. Rule-based models are classic examples of this category.

In this work, we follow the terminologies and taxonomy provided in [8], [14]. We use the term explainability and add the following four dimensions, in addition to the dimensions identified in the previous subsection, to categorize work in explainable student performance models:

- **Method:** This refers to the machine learning method used to predict student performance. Methods are categorized into broad areas: rule-learning based algorithms, decision-tree based algorithms, deep learning algorithms, and other machine-learning based algorithms.
- **Stage:** This dimension identifies the stage at which explanation occurs. There are two stages at which explanation can occur; ante-hoc and post-hoc.
  - *Ante-hoc* denotes explanations that occur during the training of a machine learning model. In other words, this refers to machine learning models that are explainable in their original format. Rule-based models are examples of ante-hoc models. In literature, these models are often denoted as white-box models, interpretable or transparent models [17].
  - *Post-hoc* denotes explanation that occurs on top of a black-box machine learning model. Figure 3 illustrates the difference between ante-hoc and post-hoc stages.
- **Scope:** This dimension refers to the scope of the explanation, where it can either be global or local.
  - *Global* denotes explanations at the model-level. That is, the explanation covers the complete model. All ante-hoc models are global, as they provide a complete explanation of the learned model.
  - *Local* denotes instance-level explainability. For example, LIME [18] is a local method of explainability which is used on top of black-box models to provide explanations at an instance level.
- **Explainable Output Type:** This dimensions refers to the explainable output. The output can either be numerical, rule-based, textual, visual or mixed.

Synthesizing studies in student performance prediction models and XAI resulted in a total of nine dimensions, which will be used to categorize primary studies in explainable student performance models. Figure 2 illustrates the taxonomy containing all dimensions.

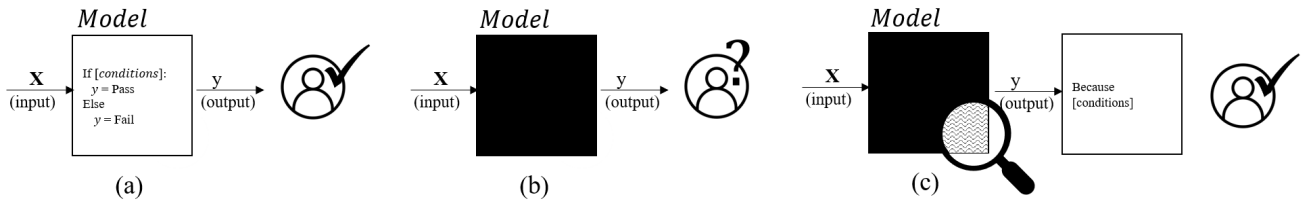


FIGURE 3. Different stages at which explanation occurs; (a) represents an ante-hoc stage, where explanation occurs during training (the learned model is explainable), (b) represents a black-box model with no explanation, (c) represents the post-hoc explanation where a model is built on top of a black-box model to explain the results.

B. RELATED WORK

In this section, we summarize secondary studies (review papers) on grade prediction models that were published in the last five years. Systematic reviews of educational data mining in general, as in [19], [20] as well as informal literature reviews as in [21] are excluded from this summary. Ideally, relevant review papers should cover *explainable* grade prediction models. However, a keyword search on two databases, Science Digital Library and Google Scholar, revealed no systematic review papers highlighting models' explainability. Consequently, discussion of related work in this section covers systematic review articles of student performance prediction models in general. Table 1 lists the search string for each data source and the number of articles identified in total, as well as per year. Additionally, forward and backward snowballing techniques were conducted in order to identify additional secondary studies.

TABLE 1. Database search strings for secondary studies.

Source	Search String
Web of Science	TI=(systematic) AND TI=(review OR survey) AND TI=(student* OR academic) AND TI=(predict* or identif*) AND TI=(performance or success or grade or at-risk or "at risk" or achievement or dropout or drop-out or "drop out")
Google Scholar	allintitle: (systematic) (review OR survey) (student OR academic) (prediction OR identification OR modelling OR detection) (performance OR success OR grade OR at-risk OR achievement OR dropout OR "early warning system")

Overall, there is a total of twelve identified systematic review articles in the last five years. Eleven of them were identified through database search, and one by backward snowballing. Nine of the eleven were identified by Web of Science and Two by Google Scholar. Figure 4a visualizes the distribution of search results per source and over years. The year 2019 has the most share of identified records, yet it is important to note that the number of identified records of 2020 might change as this search was conducted on August 2020.

The twelve identified articles are then screened, and only systematic articles that met the following criteria were selected:

- The article focuses on machine learning models
- The article is at least six pages
- The article is not a duplicate (or shorter version) of another selected article

Six out of the twelve articles, namely [22]–[27] were eliminated because they did not meet the first criteria. One article, [28], was eliminated because it did not meet the second criteria. Another article was eliminated, [28], because it was a shorter version of another article [29]. Eventually, four out of the twelve articles have met the above criteria and were considered as relevant secondary studies. The list of all secondary studies, decisions and justifications is available here.[1]

Table 2 provides a summary of eligible secondary studies, chronologically ordered. The table compares secondary studies in terms of: the focus of the review (reflected by the research question), the years covered, and the size (number of) primary studies. Reference [30] focused on identifying the *factors* and *methods* used in the higher education context, where the number of primary studies was 36. Reference [13] had the largest size of primary studies (357) where the objective was to synthesize current state of the art in predicting student performance, which covered *predictors* and *methods*. Reference [31] had 23 primary studies and focused on the main *predictors* and *methods* used in studies of at-risk student identification. Reference [29] focused on the *predictors* and *predicted value*, while highlighting the importance of big data. This study reviewed 59 primary articles.

Similar to our objective, all the above work analyzed the primary studies in terms of the utilized method. However, none of the secondary studies reviewed the primary studies in terms of the models' explainability. In this work, our objective is to focus on explainable models, where we compare and analyze primary studies that meet the inclusion criteria specified in section III. We drew upon these similar reviews when formulating our methodology, including search keywords, inclusion and exclusion criteria. Additionally, we have used the same terminologies and definitions articulated by [13].

III. MATERIAL AND METHOD

We adopted the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) methodology [32]. The PRISMA methodology aims to facilitate a transparent reporting of systematic reviews by providing a detailed checklist and a flow diagram for authors to follow. In this

[1] https://tinyurl.com/yaa54eb2

TABLE 2. Summary of secondary studies. The table summarizes relevant secondary studies in terms of the focus of the study (reflected by the research question), the years covered and the size of the study (i.e., the number of primary studies).

Ref.	Research Question (Focus of the Review)	Years Covered	Size
[30]	1. What are the factors affecting students' performance in higher education? 2. What are the data mining techniques used to analyze and predict the students' performance?	2009-2018	36
[13]	1. What is the current state of the art in predicting student performance? 1.1 How is performance defined? What types of metrics are used for describing student performance? 1.2 What are the features used for predicting performance? 1.3 What methods are used for predicting performance? 1.4 Which feature and method combinations are used to predict which types of student performance? 2. What is the quality of the work on predicting student performance?	2010-2017	357
[31]	1. The purposes of the research conducted related to at-risk students and where it was carried out 2. The types of data or attributes involved in identifying at-risk students 3. The types of tools or analytical methods used to identify at-risk students 4. The strategies or interventions suggested to help at-risk students.	2008-2017	23
[29]	1. What are the predictive models in data mining that has been used to predict students' performance? 2. What are the precise purposes of the predictive model used for within student's performance improvement context? 3. Is big data considered in student performance prediction model? 4. What are the key attributes mainly used in the predictive model and is there a selection process?	2012-2016	59

section, we describe in details the research questions and methodology.

A. RESEARCH QUESTIONS

The goal of this work is to synthesize studies that proposed explainable student performance prediction models. Specifically, we investigate the following research questions:

- **RQ1:** What are the student performance measures to be *predicted*?
- **RQ2:** What are the *predictors* used to train an explainable model?
- **RQ3:** What are the explainable machine learning *methods* used to predict students' performance?
- **RQ4:** What are the *evaluation metrics* used to assess the explainability of the models?
- **RQ5:** What are the methods that meet both requirements of high accuracy and explainability?

B. RESEARCH METHODOLOGY

The strategy we used to find relevant researches and scientific papers started with analyzing the research questions and deciding the search terms. Then, by setting the period from 2015 to 2020, we applied the search terms in Web of Science and Google Scholar. Table 3 lists the search string for the two databases, and Figure 4b visualizes the number of articles retrieved by each database, each year.

TABLE 3. Database Search strings for primary studies.

Source	Search String
Web of Science	TI=(student* OR academic) AND TI=(predict* or model*) AND TI=(performance or grade) AND TS=(white-box OR Interpretable OR understandable OR explainable OR expressiveness OR Rule*)
Google Scholar	allintitle: (student OR academic) (prediction OR model) (white-box OR interpretable OR understandable OR explainable OR expressiveness OR Rule)

1) INCLUSION AND EXCLUSION CRITERIA

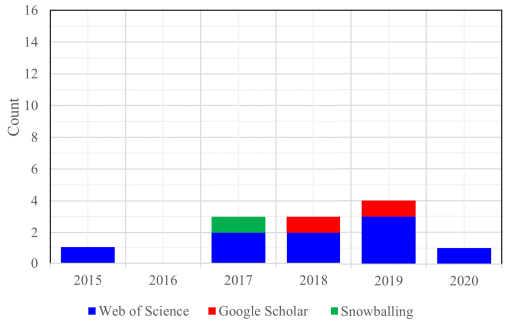
Identifying the primary studies to be included in this study is not an easy task. This is mainly because the problem of predicting students' performance in general, as well as identifying key factors affecting students' performance in specific, is multidisciplinary. It has attracted scientists from various fields, ranging from education [33] to machine learning. This lead us to define some constraints to screen eligible studies. Following is a list of the inclusion and exclusion criteria adopted in this work.

- Inclusion Criteria:
  1) Peer-reviewed articles, published between 2015 to 2020
  2) Articles on *explainable* machine-learning based student performance prediction models
- Exclusion Criteria:
  1) Not student performance prediction articles
  2) Not machine-learning based articles
  3) Machine-learning based articles with no emphasis on explainability of the model
  4) Articles not written in English

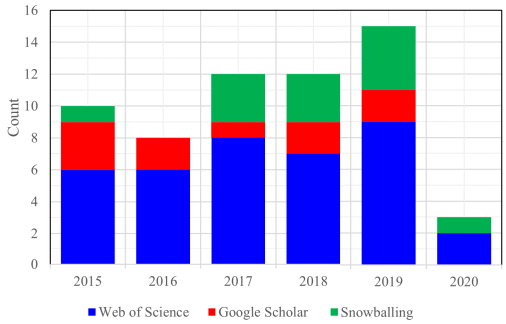
Screening was done on two stages: title/abstract screening, and full-text screening. The same inclusion/exclusion criteria were applied at these stages.

2) BACKWARD AND FORWARD SNOWBALLING

To ensure that our study captures the most number of relevant articles, we applied backward and forward snowballing [34]. This process started after we searched and screened primary



(a) Secondary studies



(b) Primary studies

FIGURE 4. Distribution of search results per data source, for secondary and primary studies.

articles. For all the eligible primary articles, we conducted a 1-iteration backward and forward snowballing. Backward snowballing refers to the approach where We check the articles' related work section and references to identify new primary articles. 1-iteration backward snowballing means that we do this process one time (for all the identified primary articles). 2-iteration backward snowballing means that we repeat this process for the second batch of identified primary studies. The reason for not doing 2-iterations, is that we did not find any new articles when we checked the references of the 1-iteration batch. Forward snowballing, on the other hand, refers to the process of checking who cited the primary articles. Similarly to backward snowballing, we performed 1-iteration forward snowballing.

3) DATA EXTRACTION

We built a data extraction form, to facilitate synthesizing primary studies. The form included all the dimensions described in Section II-A and visualized in Figure 2.

4) QUALITY ASSESSMENT

To assess the quality of our study, details of the search results and screening, is available here.[2] The form includes the following: search results and screening for secondary studies (related work), search results and screening for primary studies, and data extraction for primary articles. The forms include justification to each decision made for each article, during the title/abstract screening and full-text screening.

IV. RESULTS

The PRISMA flow diagram, detailing article identification and screening process is detailed in Figure 5. Our keyword search in the two databases resulted in a total of 48 articles, and 12 articles were identified through snowballing. The distribution of articles found per year is visualized in Figure 4b. After removing duplicates, we had a total of 56 articles identified. We then applied screening based on title and abstract only, and excluded 9 articles at this stage. Three of the articles were not about student performance prediction models, one was not machine-learning based, and the rest were not focusing on explainable machine-learning models. The initial screening resulted in a total of 47 articles to be assessed for eligibility.

Out of the 47 assessed articles, 32 were excluded with reasons. Three of them were excluded because we could not access them during the time of writing, though through our institute, we have access to most digital libraries and databases. Two of the articles were rejected because they are not machine-learning based student performance prediction models. The remaining articles (total of 27) were rejected because they do not focus on explainable machine-learning models. This resulted in a total of 15 articles that were included in the study.

Data from the eligible articles were extracted given the dimensions in Figure 2. The extracted data is available here.[3] Figures 6 to 8 visualize the distribution of data for each dimension. The context of the problem can be summarised using the following dimensions: education level (Figure 6a), performance level (Figure 6b) and problem type (Figure 6c). Distribution of the data extracted for these dimensions is visualized in Figure 6. From the figure we can conclude that most of the articles focused on higher education, with (93%), and only few articles studied the problem of explainable student performance models in K-12. Figure 6c shows that almost half of the articles aimed to solve a multi-class classification problem. This was followed by binary classification, where 30% of the articles treated the student performance prediction problem as a binary classifier. Only 13% of the reviewed articles tackled the student performance prediction as a regression problem. For example, studies that predict whether an undergraduate student will fail or pass a course, a.k.a., at risk students, is categorized as educational level - higher education, performance level - course level, and problem type - binary classification. Similarly, studies that predicts the letter grade at the end of an undergraduate course is categorized as educational level - higher education, performance level - course level, and problem type - multi-class classification.

The next two dimensions provide further technical details on the input to the model. This includes predictors and

[2]https://tinyurl.com/yaa54eb2

[3]https://tinyurl.com/yaa54eb2

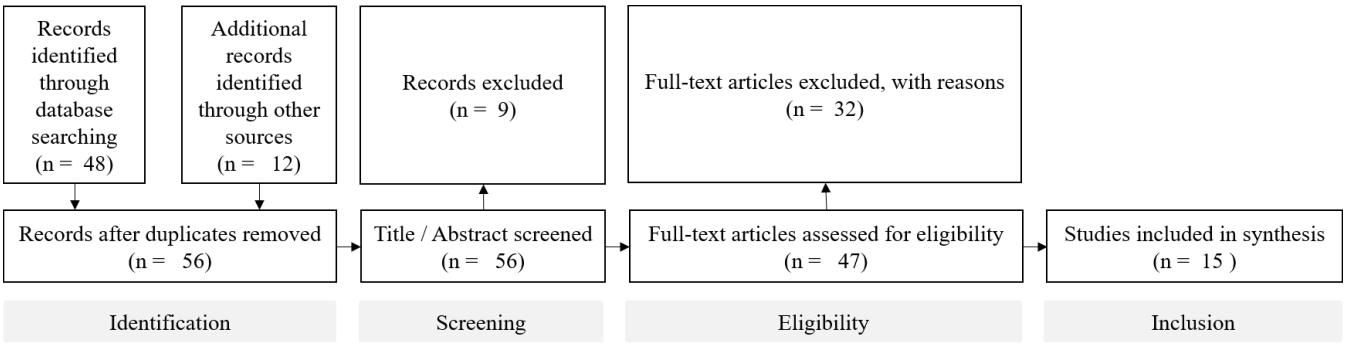
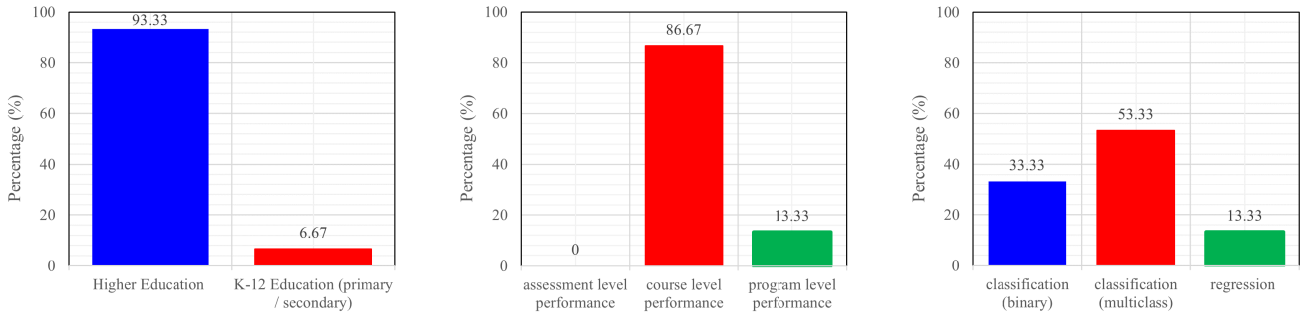
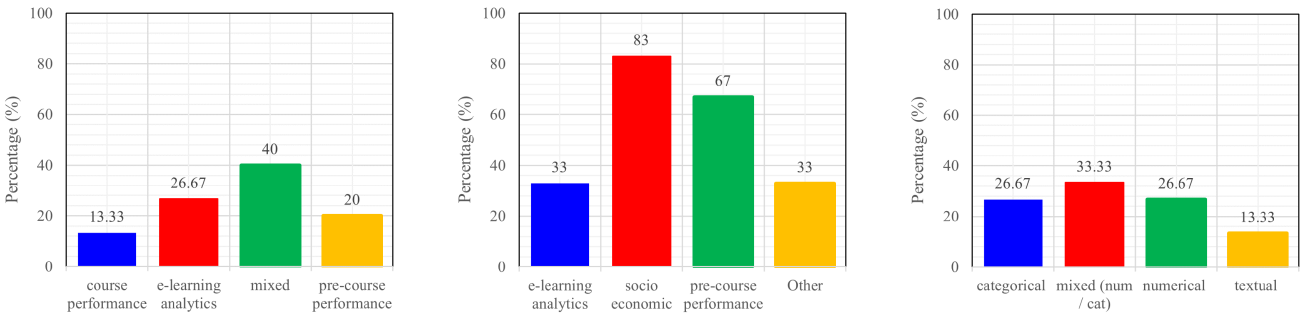


FIGURE 5. PRISMA flow diagram for systematic search and study selections.



(a) Educational Level (b) Performance Level (c) Problem Type

FIGURE 6. Distribution of primary studies by the general problem context dimensions. This includes educational level, performance level, and problem type.

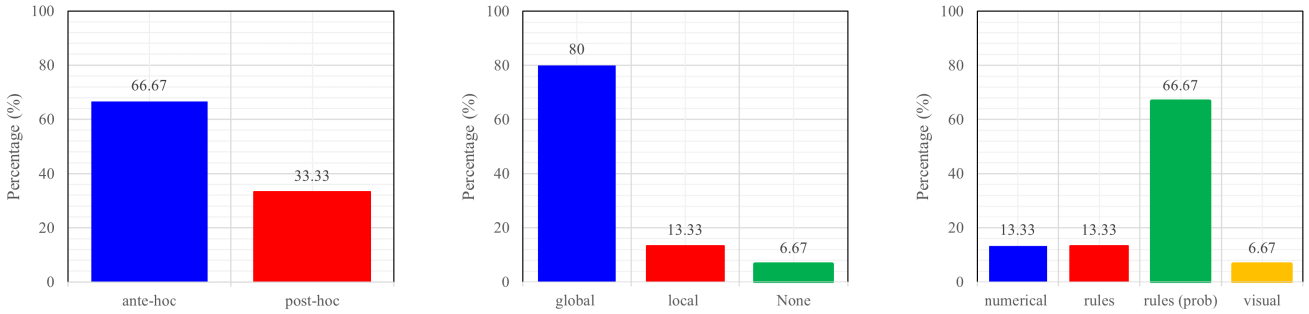


(a) Predictors (b) Mixed Predictors (c) Predictors Type

FIGURE 7. Distribution of primary studies by the input to the model. This includes predictors and predictors' type.

predictor type. Figure 7 visualizes the distribution of data for each dimension. Figure 7a illustrates that the majority of the primary studies (40%) uses a combination of mixed predictors. Mixed predictors means that the study used more than one data source to predict student performance. The majority of these studies utilized socio-economic data (80%) along with pre-course performance measures (67%). Additionally, 30% of the studies included e-learning analytic and other data sources as well. Figure 7b visualizes the distribution of different predictors used in studies marked as mixed predictors. For studies that only included one predictor, e-learning analytic was the most common predictor (26%), followed by pre-course performance (20%) and course performance data (13%). As for predictor type, in Figure 7c, the majority of papers used both numerical and categorical predictors (33%), this is followed by 26% for both categorical and numerical types. Lastly, only 13% of the primary studies utilized textual data for student performance prediction.

Table 4 categorizes the context of the primary studies, in light of the general problem dimensions. The study by [35] used a collection of e-learning data, pre-course data and socio-economics to predict at-risk students (pass/fail). In this



(a) Stage (b) Scope (c) Explained Output

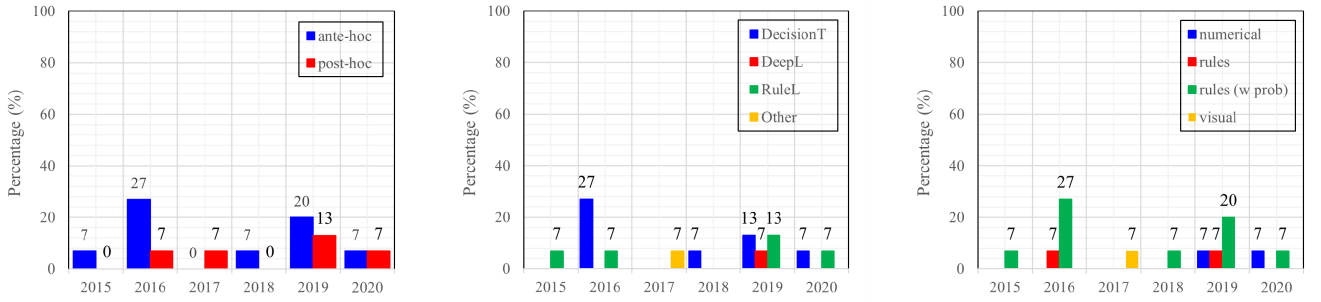
FIGURE 8. Distribution of primary studies by the stage, scope and output of the explainable model.

TABLE 4. Categorization of primary studies in terms of the five student performance model dimensions; i.e., problem type, educational level, performance level, predictors and predictors' type.

Problem type	Educational Level	performance level	Predictors	Predictors' Type	Ref
Classification (binary)	Higher Education	Course level	Mixed	Categorical	[35]
				Mixed (numerical / categorical)	[36], [37]
			Pre-course performance	Categorical	[38]
		Program level	Mixed	Mixed (numerical / categorical)	[39]
Classification (multiclass)	Higher Education	Course level	Course performance	Textual	[40], [49]
				Numerical	[41], [42], [43]
			Mixed	Categorical	[44]
		Program level	Pre-course performance	Categorical	[45]
	K-12	Course level	Mixed	Mixed (numerical / categorical)	[46]
Regression	Higher Education	Course level	e-learning analytics	Numerical	[47]
			Pre-course performance	Numerical	[48]

work, some of the predictors (e.g., gender) are categorical, while others are originally numerical but were categorized (discretized) in the pre-processing phase. All numerical predictors were categorized into four quartiles. Similarly, the studies by [36] and [37] built an early warning system to identify at-risk students. Unlike the previous study, the authors utilize both numerical and categorical predictors' type. The study by [38] built a model that can predict student performance in a course as either good or bad, given their performance in previous courses. Performance in previous courses was categorized as either good or bad, where good indicates a B grade or higher, and bad indicate a C+ grade or lower. Additionally, the study by [39] builds an early detection system of at-risk students. In this study, students are predicted to either have satisfactory or poor performance, and prediction was based on socio-economic data and pre-course data. All these studies have tackled the student performance prediction problem as a binary classification task.

The majority of primary studies, on the other hand, have tackled this problem as a multi-class classification problem. In this setting, the model predicts one of three or more outputs. For example, the study by [40] predicts student performance in a four-way scale (A, B, C and D). The study utilizes student comments collected after each lesson to build the classifier. Similarly, the work by [41] predicts student performance in a five-way scale; excellent, good, average, sufficient, and at-risk, where the input to the model are numerical attributes. In the study by [42], course-level performance is categorized into four-way scale, which are: perfect, excellent, good and sufficient. The work by [43] categorizes course-level performance as one of three classes: high performance, medium performance and low performance. The study by [44] predicts conventional letter grades from C up to A+. The predictors were a mix of socio-economic data and pre-course data, and the predictors' type was categorical. Additionally, in multi-class classification, one primary paper has studied the performance at a program-level, using categorical pre-course data [45]. Another study by [46] focused on course-level performance on K-12 grades using socio-economic and environmental data.



(a) Stage over time (b) Method over time (c) Explainable output over time

FIGURE 9. Distribution of primary studies, over year of publication, by the stage, method and output of the explainable model.

TABLE 5. Categorization of primary studies in terms of the four explainable model dimensions; i.e., stage, scope, method, explainable output type.

Problem type	Stage	Scope	Method	Explainable output type	Ref
Classification (binary)	Ante-hoc	global	Decision Tree learning algorithms	Rules (w prob)	[35]
			Rule learning algorithms	Rules	[36]
				Rules (w prob)	[37], [38]
	Post-hoc	Local	Decision Tree learning algorithms	Numerical	[39]
Classification (multiclass)	Ante-hoc	Global	Decision Tree learning algorithms	Rules	[45]
				Rules (w prob)	[46], [40]
			Rule learning algorithms	Rules (w prob)	[42], [41]
	Post-hoc	Global	Decision Tree learning algorithms	Rules (w prob)	[44], [49]
		None	Other Machine learning algorithms	visual	[43]
Regression	Ante-hoc	Global	Decision Tree learning algorithms	Rules (w prob)	[47]
	Post-hoc	Local	Deep Learning	Numerical	[48]

A third group of primary studies focused on predicting numerical grades. Both studies in this category proposed models for course-level student performance prediction in a higher education setting. Both studies used numerical input, where one utilized e-learning analytic [47] and the other one [48] used pre-course performance data for prediction.

The remaining dimensions focus on the explainable machine learning models employed. This includes: method, stage, scope and explainable output, as defined in section II-A. In general, the methods used in order are: decision tree algorithms (53%), rule learning based algorithms (33%), deep learning algorithms (6%) and other machine learning based algorithms (6%). Figure 8 illustrates summary statistics on stage, scope and explained output. The majority of models in this field adopt the ante-hoc approach, where the learned model itself is explainable. Additionally, 80% of the studies adopts global scope, where explanations occur at the model level and not the instance-level. Moreover, the majority of studies (66%) adopted the rule based explainable output with probabilities. In this approach, the output of the model is a list of rules with probabilities, that the model uses to make predictions. Figure 9 visualizes the distribution of the three metrics (stage, method and explainable output) over time. From the distributions, we can observe the following. Ante-hoc models are more commonly used than post-hoc models, yet the gap in usage decreased over time (observe the gap in year 2016 and 2019). This might indicate that an increase in the usage of post-hoc models might be observed in the coming few years. This is especially true if larger datasets are used, which require deeper and more complex models. In this case, adopting a post-hoc approach on top of a black-box model can achieve the goal of explainability. This observation is supported by Figure 9b where we observe adopting deep learning models in the later years (2019), and the peek for decision tree models was in 2016, and decreased after that. The distribution of explainable output is almost uniform over time, except with two peeks caused by rule based output with probabilities.

Table 5 summarizes the context of the primary studies, in light of the explainable models dimensions. It is interesting to observe the relationship between stage, method, and explained output, given the primary studies. Figure 10 depicts this observed relationship. The thickness of lines denote the number of time a primary study followed this approach. For example, the figure shows that all rule learning based algorithms were ante-hoc. This means that all rule based algorithms are explainable in nature. This does not apply to all decision tree algorithms. A decision tree algorithm can be ante-hoc, as in [45], where the learned decision tree is explainable. It can also be post-hoc, as in [44], where

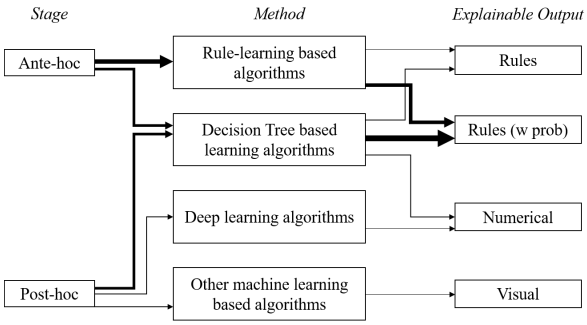


FIGURE 10. Visualization of the relationships between stage, method and explainable output. The thickness of lines is relative to the number of time a primary paper used this approach.

the learned tree is too complex, and it requires another model on top of it to extract rules from the tree. It is also interesting to observe that the most commonly used explainable output, in this setting, is rules with probability. This is justifiable, as rules are easy to interpret and understand, and probability provide another layer of confidence to the model.

Given the above analysis, we now answer the research questions listed in Section III-A.

RQ1: What are the student performance measures to be *predicted*?

There are a variety of student performance measures to be predicted. In general, student performance models can be grouped by the problem type, educational level and performance level. As our analysis shows, multi-class classification is most common in this area of research, followed by binary classification, then regression. The majority of studies focused on higher education context, and course level performance. This means that the most common student performance measures to be predicted are expected categorical performance of a certain course, in a higher educational institute.

RQ2: What are the *predictors* used to train an explainable model?

Mostly, a combination of socio-economic features and pre-course performance features. Although the use of e-learning analytic has increased over the past few years, in student performance prediction. This data source of e-learning analytic is not fully utilized in explainable student performance models.

RQ3: What are the explainable machine learning *methods* used to predict students' performance?

The most commonly used algorithms are: decision tree and rule based learning algorithms. Table 6 lists the methods used in all the general method category. It is important to note that deep learning models are being utilized in student performance prediction. However, the explainability of these models are not being considered. This highlights an important research gap that needs to be explored and investigated thoroughly.

RQ4: What are the *evaluation metrics* used to assess the explainability of the models?

TABLE 6. Categorization of primary studies in terms of the method used.

Method (General)	Method Name
Decision Tree learning algorithms	CART [39], [46]
	J48 [45], [35]
	Jrip [35]
	Random Forest [40], [44], [49], [39]
	unspecified [47]
Deep Learning	LSTM [48]
Other Machine learning algorithms	SVM [49], [39]
	RBF [43]
	Logit [39]
	NB [39]
Rule learning algorithms	CN2 rule inducer algorithm [37]
	Classification association rule mining [38]
	Genetic-based algorithms [41], [36], [42]

None of the primary studies that were included in this systematic literature review has utilized any evaluation metric to assess the explainability of the model. This reveals a critical shortcoming to existing research in explainable student performance prediction models. For more on explainability evaluation metrics, we refer the reader to [50] and]citecarvalho2019machine.

RQ5: What are the methods that meet both requirements of high accuracy and explainability?

Because none of the studies used a quantifiable metric to measure explainability, this question could not be answered.

V. THREATS TO VALIDITY

The systematic review described in this work followed the PRISMA methodology [32], to ensure comprehensive reporting of identifying, selecting and critically evaluating relevant work. Coverage of all relevant work is one potential threat to validity for this review. We have used a comprehensive set of keywords to minimize the number of left out relevant work. We have also searched the two key data sources in this field, which are Web of Science and Google Scholar [51]. Additionally, we did backward and forward snowballing to incorporate any work that might have not been included either because of the search terms or data sources.

A second threat to validity includes our explicit use of ''interpretable'' related words in the search string. This is justified because our objective is to capture all work that utilizes either interpretable or explainable models. However, we are aware of many relevant work that utilizes interpretable models, such as decision trees or rule-based models, without explicitly using the terms interpretable or white-box models. Examples of such work include [52]. This being said, it is important to note here that we are mainly interested in papers that used such models, and evaluated the interpretability of these models, in which case such terms would appear in the paper.

Another possible threat to validity is article screening and eligibility assessment. This process was conducted independently by the two authors of this work, where the predefined

inclusion and exclusion criteria were followed. The result of the screening and eligibility assessment can be viewed here.[4]

Finally, the analysis derived from this systematic review applies to white and black-box models applied in the student performance prediction context, and cannot be generalized to other domains or fields.

VI. CONCLUSION

The main objective of this systematic review is to identify state-of-the-art in explainable student performance models. To achieve this objective, we first defined five research questions which highlighted four main aspects of the models. These aspects are: the student performance measure to be predicted (the output), the predictors used (the input), the explainable methods used (the model), and the evaluation metrics used to assess the performance of the models. Then, we systematically reviewed the literature and synthesized existing work from the past five years. The results of our synthesis revealed that most of the studies in explainable student performance models focused on predicting student outcomes per course, usually as a multi-class problem. The analysis also revealed that socio-economic features and pre-course performance are the top predictors used in current studies. Additionally, decision trees and rule based learning algorithms were the common machine learning methods used in such studies.

These findings highlight the gaps in research in the area of explainable student performance models. State of the art explainable models have been utilized in many domains [17]. However, they are yet to be explored in educational data mining. In the context of explainable student performance models, there is a lack of studies that adapt state of the art explainable methods and utilize rich predictors such as e-learning analytics to predict student performance in different levels. Additionally, another key limitation is concerned with the lack of adopting evaluation metrics for model explainability. The results of our synthesis revealed that none of the existing studies have utilized evaluation metrics to assess the explainability of the proposed models. This makes it difficult to compare explainability level of different models. In such studies, assessing the explainability level is as important as evaluating the prediction accuracy of the model. This current limitation can be overcame by investigating state of the art metrics in this area and use them to evaluate current models. To conclude, this study sheds light on the importance of explainable models in the educational setting, and highlights main limitations in existing literature and potential future work.

ACKNOWLEDGMENT

The authors acknowledge the technical and financial support of University of Jeddah.

[4]https://tinyurl.com/yaa54eb2

REFERENCES

[1] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Comput. Educ.*, vol. 143, Jan. 2020, Art. no. 103676.
[2] F. Yang and F. W. B. Li, "Study on student performance estimation, student progress analysis, and student potential prediction based on data mining," *Comput. Educ.*, vol. 123, pp. 97–108, Aug. 2018.
[3] B.-H. Kim, E. Vizitei, and V. Ganapathi, "GritNet: Student performance prediction with deep learning," 2018, *arXiv:1804.07405*. [Online]. Available: http://arxiv.org/abs/1804.07405
[4] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading," *Comput. Educ.*, vol. 103, pp. 1–15, Dec. 2016.
[5] M. Kumar, A. J. Singh, and D. Handa, "Literature survey on educational dropout prediction," *Int. J. Educ. Manage. Eng.*, vol. 7, no. 2, pp. 8–19, Mar. 2017.
[6] X. Du, J. Yang, and J.-L. Hung, "An integrated framework based on latent variational autoencoder for providing early warning of at-risk students," *IEEE Access*, vol. 8, pp. 10110–10122, 2020.
[7] J. Xiao, M. Wang, B. Jiang, and J. Li, "A personalized recommendation system with combinational algorithm for online learning," *J. Ambient Intell. Humanized Comput.*, vol. 9, no. 3, pp. 667–677, Jun. 2018.
[8] R. Francesca, "Ai ethics for enterprise AI," IBM Res., Tech. Rep., 2019. [Online]. Available: https://economics.harvard.edu/files/economics/files/rossi-francesca_4-22-19_ai-ethics-for-enterprise-ai_ec3118-hbs.pdf
[9] A. Elbadrawy and G. Karypis, "Domain-aware grade prediction and Top-n course recommendation," in *Proc. 10th ACM Conf. Recommender Syst.*, Sep. 2016, pp. 183–190.
[10] E. L. David, *The Prediction of Academic Performance*. New York, NY, USA: Russel Sage Found., 1965.
[11] M. Rahman Minar and J. Naher, "Recent advances in deep learning: An overview," 2018, *arXiv:1807.08169*. [Online]. Available: http://arxiv.org/abs/1807.08169
[12] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, and A. Holzinger, "Explainable artificial intelligence: Concepts, applications, research challenges and visions," in *Proc. Int. Cross-Domain Conf. Mach. Learn. Knowl. Extraction*. Cham, Switzerland: Springer, 2020, pp. 1–16.
[13] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom, and S. N. Liao, "Predicting academic performance: A systematic literature review," in *Proc. Companion 23rd Annu. ACM Conf. Innov. Technol. Comput. Sci. Educ.*, Jul. 2018, pp. 175–199.
[14] G. Vilone and L. Longo, "Explainable artificial intelligence: A systematic review," 2020, *arXiv:2006.00093*. [Online]. Available: http://arxiv.org/abs/2006.00093
[15] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.
[16] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[17] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019.
[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.
[19] X. Du, J. Yang, J.-L. Hung, and B. Shelton, "Educational data mining: A systematic review of research and emerging trends," *Inf. Discovery Del.*, vol. 48, no. 4, pp. 225–236, May 2020.
[20] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15991–16005, 2017.
[21] S. K. Mohamad and Z. Tasir, "Educational data mining: A review," *Procedia Social Behav. Sci.*, vol. 97, pp. 320–324, Nov. 2013.
[22] M. Wolden, B. Hill, and S. Voorhees, "Predicting success for student physical therapists on the national physical therapy examination: Systematic review and meta-analysis," *Phys. Therapy*, vol. 100, no. 1, pp. 73–89, 2020.
[23] T. Goradia and A. Bugarcic, "Exploration and evaluation of the tools used to identify first year at-risk students in health science courses: A systematic review," *Adv. Integrative Med.*, vol. 6, no. 4, pp. 143–150, Dec. 2019.

[24] K. Mthimunye and F. M. Daniels, "Predictors of academic performance, success and retention amongst undergraduate nursing students: A systematic review," *South Afr. J. Higher Educ.*, vol. 33, no. 1, pp. 200–220, Apr. 2019.
[25] P. J. A. C. van der Zanden, E. Denessen, A. H. N. Cillessen, and P. C. Meijer, "Domains and predictors of first-year student success: A systematic review," *Educ. Res. Rev.*, vol. 23, pp. 57–77, Feb. 2018.
[26] G. Crisp, A. Taggart, and A. Nora, "Undergraduate Latina/o students: A systematic review of research identifying factors contributing to academic success outcomes," *Rev. Educ. Res.*, vol. 85, no. 2, pp. 249–274, 2015.
[27] B. C. G. Costa and D. D. S. Fleith, "Prediction of academic achievement by cognitive and socio-emotional variables: A systematic review of literature," *Trends Psychol.*, vol. 27, no. 4, pp. 977–991, 2019.
[28] X. Hu, C. W. L. Cheong, W. Ding, and M. Woo, "A systematic review of studies on predicting student learning outcomes using learning analytics," in *Proc. 7th Int. Learn. Anal. Knowl. Conf.*, Mar. 2017, pp. 528–529.
[29] S. Muthukrishnan, M. Govindasamy, and M. Mustapha, "Systematic mapping review on student's performance analysis using big data predictive model," *J. Fundam. Appl. Sci.*, vol. 9, no. 4S, pp. 730–758, 2017.
[30] A. A. Saa, M. Al-Emran, and K. Shaalan, "Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques," *Technol., Knowl. Learn.*, vol. 24, no. 4, pp. 567–598, 2019.
[31] K. S. Na and Z. Tasir, "Identifying at-risk students in online learning by analysing learning behaviour: A systematic review," in *Proc. IEEE Conf. Big Data Analytics (ICBDA)*, Nov. 2017, pp. 118–123.
[32] D. Moher, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Ann. Internal Med.*, vol. 151, no. 4, p. 264, Aug. 2009.
[33] H. M. Vo, C. Zhu, and N. A. Diep, "The effect of blended learning on student performance at course-level in higher education: A meta-analysis," *Stud. Educ. Eval.*, vol. 53, pp. 17–28, Jun. 2017.
[34] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proc. 18th Int. Conf. Eval. Assessment Softw. Eng. (EASE)*, 2014, pp. 1–10.
[35] S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, D. J. Murray, and Q. Long, "Predicting academic performance by considering student heterogeneity," *Knowl.-Based Syst.*, vol. 161, pp. 134–146, Dec. 2018.
[36] A. Cano and J. D. Leonard, "Interpretable multiview early warning system adapted to underrepresented student populations," *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 198–211, Apr. 2019.
[37] R. Hasan, S. Palaniappan, S. Mahmood, A. Abbas, K. U. Sarker, and M. U. Sattar, "Predicting student performance in higher educational institutions using video learning analytics and data mining techniques," *Appl. Sci.*, vol. 10, no. 11, p. 3894, Jun. 2020.
[38] G. Badr, A. Algobail, H. Almutairi, and M. Almutery, "Predicting students' performance in university courses: A case study and tool in KSU mathematics department," *Procedia Comput. Sci.*, vol. 82, pp. 80–89, 2016.
[39] P. Kumar and M. Sharma, "Predicting Academic performance of international students using machine learning techniques and human interpretable explanations using LIME—Case study of an Indian University," in *Proc. Int. Conf. Innov. Comput. Commun.* Singapore: Springer, 2020, pp. 289–303.
[40] S. E. Sorour and T. Mine, "Building an interpretable model of predicting student performance using comment data mining," in *Proc. 5th IIAI Int. Congr. Adv. Appl. Informat. (IIAI-AAI)*, Jul. 2016, pp. 285–291.
[41] W. Xing, R. Guo, E. Petakovic, and S. Goggins, "Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory," *Comput. Hum. Behav.*, vol. 47, pp. 168–181, Jun. 2015.
[42] W. Zhang, Y. Zhou, and B. Yi, "An interpretable online learner's performance prediction model based on learning analytics," in *Proc. 11th Int. Conf. Educ. Technol. Comput.*, Oct. 2019, pp. 148–154.
[43] C. J. Villagrá-Arnedo, F. J. Gallego-Durán, F. Llorens-Largo, P. Compañ-Rosique, R. Satorre-Cuerda, and R. Molina-Carmona, "Improving the expressiveness of black-box models for predicting student performance," *Comput. Hum. Behav.*, vol. 72, pp. 621–631, Jul. 2017.
[44] E. C. Abana, "A decision tree approach for predicting student grades in research project using weka," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, 2019.
[45] M. A. Al-Barrak and M. Al-Razgan, "Predicting students final GPA using decision trees: A case study," *Int. J. Inf. Educ. Technol.*, vol. 6, no. 7, p. 528, 2016.
[46] I. Meca, N. Mollá-Campello, and A. Rabasa, "A new methodology for early warning of critical academic performance, based on discrete predictive models," in *Proc. 7th Int. Conf. Technol. Ecosyst. Enhancing Multiculturality*, Oct. 2019, pp. 680–685.
[47] A. Pardo, N. Mirriahi, R. Martinez-Maldonado, J. Jovanovic, S. Dawson, and D. Gašević, "Generating actionable predictive models of academic performance," in *Proc. 6th Int. Conf. Learn. Analytics Knowl. (LAK)*, 2016, pp. 474–478.
[48] Q. Hu and H. Rangwala, "Reliable deep grade prediction with uncertainty estimation," in *Proc. 9th Int. Conf. Learn. Analytics Knowl.*, Mar. 2019, pp. 76–85.
[49] S. E. Sorour, S. A. E. Rahman, S. A. Kahouf, and T. Mine, "Understandable prediction models of student performance using an attribute dictionary," in *Proc. Int. Conf. Web-Based Learn.* Cham, Switzerland: Springer, 2016, pp. 161–171.
[50] A.-P. Nguyen and M. Rodríguez Martínez, "On quantitative aspects of model interpretability," 2020, *arXiv:2007.07584*. [Online]. Available: http://arxiv.org/abs/2007.07584
[51] J. C. Fagan, "An evidence-based review of academic Web search engines, 2014-2016: Implications for librarians' practice and research agenda," *Inf. Technol. Libraries*, vol. 36, no. 2, pp. 7–47, 2017.
[52] R. Hasan, S. Palaniappan, A. R. A. Raziff, S. Mahmood, and K. U. Sarker, "Student academic performance prediction by using decision tree algorithm," in *Proc. 4th Int. Conf. Comput. Inf. Sci. (ICCOINS)*, Aug. 2018, pp. 1–5.

RAHAF ALAMRI received the B.Sc. degree in information systems from the University of Jeddah, Jeddah, Saudi Arabia, in 2019. She is currently working as a Researcher with various faculty members with the University of Jeddah. Her research interests include data mining and applied machine learning.

BASMA ALHARBI received the B.Sc. degree in computer science from Effat University, Jeddah, Saudi Arabia, in 2008, the M.Sc. degree in computer science from Durham University, Durham, U.K., in 2009, and the Ph.D. degree in computer science from the King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, in 2017. She is currently an Assistant Professor with the Computer Science and AI Department, College of Computer Science and Engineering, University of Jeddah, Jeddah.