



E-learningDJUST: E-learning dataset from Jordan university of science and technology toward investigating the impact of COVID-19 pandemic on education

Malak Abdullah¹ · Mahmoud Al-Ayyoub¹ · Saif AlRawashdeh¹ · Farah Shatnawi¹

Received: 3 September 2021 / Accepted: 27 October 2021 / Published online: 13 November 2021
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2021

Abstract

Recently, the COVID-19 pandemic has triggered different behaviors in education, especially during the lockdown, to contain the virus outbreak in the world. As a result, educational institutions worldwide are currently using online learning platforms to maintain their education presence. This research paper introduces and examines a dataset, E-LearningDJUST, that represents a sample of the student's study progress during the pandemic at Jordan University of Science and Technology (JUST). The dataset depicts a sample of the university's students as it includes 9,246 students from 11 faculties taking four courses in spring 2020, summer 2020, and fall 2021 semesters. To the best of our knowledge, it is the first collected dataset that reflects the students' study progress within a Jordanian institute using e-learning system records. One of this work's key findings is observing a high correlation between e-learning events and the final grades out of 100. Thus, the E-LearningDJUST dataset has been experimented with two robust machine learning models (Random Forest and XGBoost) and one simple deep learning model (Feed Forward Neural Network) to predict students' performances. Using RMSE as the primary evaluation criteria, the RMSE values range between 7 and 17. Among the other main findings, the application of feature selection with the random forest leads to better prediction results for all courses as the RMSE difference ranges between (0–0.20). Finally, a comparison study examined students' grades before and after the Coronavirus pandemic to understand how it impacted their grades. A high success rate has been observed during the pandemic compared to what it was before, and this is expected because the exams were online. However, the proportion of students with high marks remained similar to that of pre-pandemic courses.

Keywords E-learning · Machine learning · Correlation · COVID-19

Malak Abdullah and Mahmoud Al-Ayyoub have contributed equally to this work.

✉ Malak Abdullah
mabdullah@just.edu.jo

Mahmoud Al-Ayyoub
maalshbool@just.edu.jo

Saif AlRawashdeh
rawashdehsaif5@gmail.com

Farah Shatnawi
farahshatnawi620@gmail.com

¹ Computer Science, Jordan University of Science and Technology, Irbid 22110, Jordan

1 Introduction

The e-learning system helps students and teachers and enables them to continue the learning process at all available times through digital resources. It differs from traditional learning in that students and teachers are not required to be in the same classroom during the teaching process [1, 2]. Several standard terms are used interchangeably for e-Learning, like distance learning and online learning [1]. Many academic institutions have been using the e-learning system for storing, sharing, and retrieving courses materials [1, 3]. It also allows joining a vast number of students into the same course without worrying about the classroom spaces [1, 3, 4]. Assuredly, e-learning systems help educational institutes worldwide to maintain their education presence during the COVID-19 pandemic, especially during the lockdown to contain the virus outbreak in different countries.

E-learning is an important and exciting research topic that has attracted many researchers in the past few years [2, 3, 5]. This research paper proposes a new dataset, E-LearningDJUST, and examines students' study progress within Jordan University of Science and Technology (JUST). E-LearningDJUST stands for E-Learning Dataset collected from Jordan University of Science and Technology. It is obtained from the Center for E-Learning and Open Educational Resources and the Admission and Registration Unit at JUST. This dataset is collected during the spring 2020, summer 2020, and fall 2021 semesters. The collected data contains the university's students' log files of four subjects in three successive semesters for 9,246 students. The students' identity numbers and names had been masked after obtaining the required IRB to maintain the confidentiality of student information. To our knowledge, this dataset is the first dataset to be collected to reflect the student's academic progress within a Jordanian institute using the e-learning system database. This study extracts and analyzes the features from logs' events and weeks. Then, it explores the correlation between the features and the total grades (out of 100). It is worth mentioning that the essential features that impact students' grades are related to quizzes events in most of the subjects.

This study aims to explore the correlation between e-learning events and the final grades out of 100. Consequently, to help to determine the events that can give indications to students progress. Building a model that can predict students' performances based on their records helps students know and work on their weak subjects. It also helps the faculty and the parents to get alert and take appropriate measures. The current study applies three regression models to the data set to predict students' performances; two robust machine learning models (RF and

XGB) and Feed Forward Neural Networks. The results show that the systems can predict students' grades out of 100 with an RMSE range between 7 and 17. The study also conducts a statistical experiment to investigate how the Coronavirus affects students' performances. Therefore, students' grades for the four subjects were obtained for seven consecutive semesters; Four of them before the pandemic and three after it. Accordingly, this article examined the number of students who passed and failed each course in the seven semesters. It also illustrates the number of students in passing levels (90-100, 80-89, 70-79, 60-69, 50-60) for all courses. It was noted that success rates were increased in all subjects with online learning during the pandemic.

The main contribution of this article can be summarized as follows:

- Providing an e-learning dataset, E-LearningDJUST, to address the lack of dataset availability.
- Building baseline regression models to predict students' performances using e-learning records
- Exploring the correlation between e-learning use and students' performances
- Investigating the impact of coronavirus pandemic on students' performances.

The remaining sections of this paper are organized as follows: Sect. 2 presents the literature review. Section 3 describes the dataset collected and the masking and merging techniques for the dataset. Section 4 experiments with three models to predict students' performances. Section 5 studies the impact of the Coronavirus on students' performances. Finally, Sect. 6 concludes this research paper.

2 Literature review

Machine learning (ML) is a branch of artificial intelligence that uses historical data to predict outcomes without being explicitly programmed. Using machine learning allows organizations to study different patterns in customer behavior or the attitudes and opinions of various segments of people. Machine learning techniques are attracting substantial interest from many sectors. For example, in the medical sector, the authors [6] developed predictive models for cancer diagnosis using descriptions of nuclei sampled from breast masses. Another group of researchers [7] used machine and deep learning to identify and segment pneumothorax in x-ray images. Moreover, machine learning has also been used in intelligent transportation systems to predict urban traffic crowd flows. The group of researchers in [8–10] proposed a deep hybrid Spatio-temporal dynamic neural network to predict both inflows and outflows in

every region of a city. Also, the authors in [11] developed a machine learning model for recommending the most appropriate transport mode for different users. There are many essential machine learning applications in the education sector, such as predictive analytics in education for identifying the mindset and demands of the students. It helps predict which students will perform well in the exam. Also, by learning analytics, the teacher can gain insight into data making connections and conclusions to impact the teaching and learning process positively [12, 13].

Several researchers from different countries around the world had studied e-learning systems' data [14–17]. The researchers in [14] study the effect of parents' participation in the learning process. This category of features is concerned with the learner's interaction with the e-learning management system. Three different classifiers such as Naive Bayes, Decision Tree, and Artificial Neural networks were used to examine the effect of these features on students' educational performance. The accuracy of the proposed model achieved up to 10% to 15% and is much improved as compared to the results when such features are removed. Iatrellis et al. [15] collected the students' dataset from the Computer Science Department of the University of Thessaly (UTH), in Greece. This dataset included 1,100 records for graduates. Each student has 13 variables: GPA, specialization field, capstone project grade, etc. The researchers used two machine learning models to be trained on three coherent clusters of students who were grouped based on the similarity of specific education-related factors and metrics in order to predict the time to degree completion and student enrollment in the offered educational programs. Moreover, the authors in [18] obtained educational students' from e-learning at Eindhoven University of Technology (TU/e), Netherlands, of 2014/2015. This dataset contains 4,989 students from 17 courses. They aimed to predict student performance from LMS predictor variables using both multi-level and standard regressions. The analyses show that the results of predictive modeling varied across courses.

However, few researchers had studied this subject in the Middle East region [19–22]. The authors in [19] collected students' datasets from the College of Computer Science and Information Technology, University of Basrah. This dataset was collected using two questionnaires: Google Forms and an open-source application (LimeSurvey), in which the total number of questionnaires is 161. The survey consists of 60 questions that cover the fields, such as health, social activity, relationships, and academic performance, most related to and affect the performance of students. The authors built a model based on decision tree algorithms using three classifiers (J48, Random Tree, and REPTree). They found that the J48 algorithm was considered as the best algorithm based on its performance

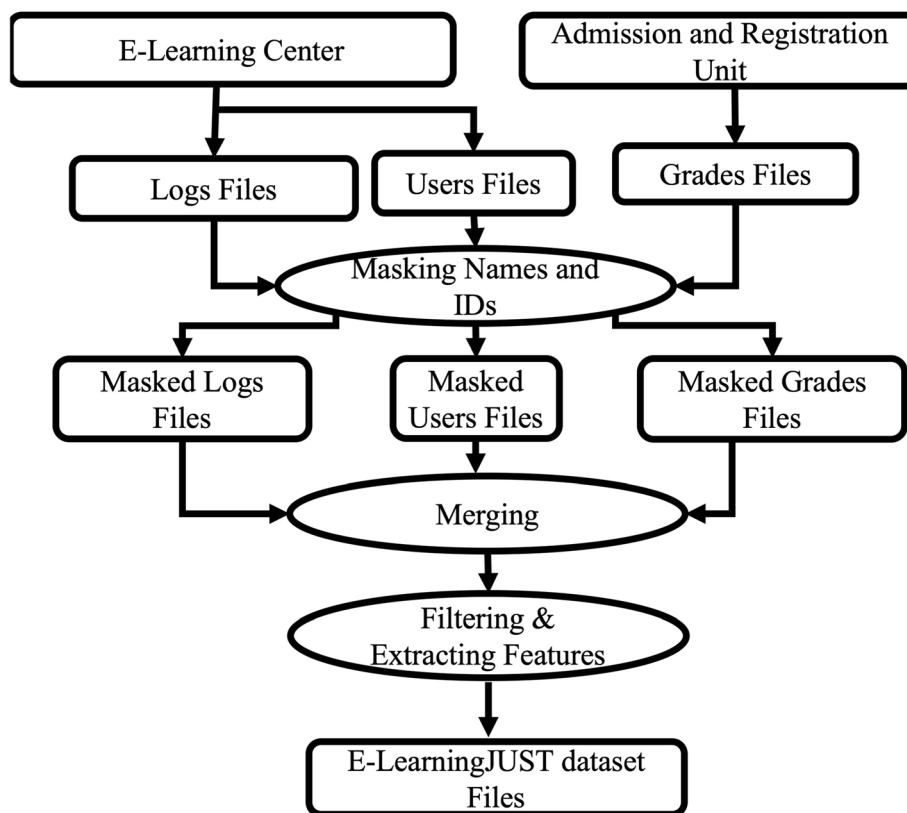
compared with the Random Tree and RepTree algorithms. Hashim et al. [20] collected students' datasets from the College of Computer Science and Information Technology (CSIT), University of Basra, Iraq, for two academic years, 2017–2018, and 2018–2019. The size of the dataset is 499 students with ten attributes. They compared the performances of several supervised machine learning algorithms to predict student performance on final examinations. Results indicated that the logistic regression classifier is the most accurate in predicting the exact final grades of students (68.7% for passed and 88.8% for failed). Zeineddine et al. [22] collected student datasets from the Admission, Registrar, and Student Service offices in the United Arab Emirates. The dataset's size is 1,491 students, including 13 features in each record for students. They proposed the use of Automated Machine Learning to enhance the accuracy of predicting student performance using data available prior to the start of the academic program.

To the best of our knowledge, no study has been conducted to examine students' study progress from Jordan's e-learning systems. This research fills this gap by introducing the E-LearningDJUST dataset in Jordan for Spring and Summer 2019/2020 semesters and the Fall 2020/2021 semester. It also explores the high correlated features with students' grades and studies the impact of the essential features on models' predictions.

3 Dataset collection

Figure 1 shows the workflow of building the E-LearningDJUST dataset corpus. The data has been collected from two units at Jordan University of Science and Technology for three consecutive semesters (Spring/2020, Summer/2020, Fall/ 2021). First, we have selected the courses with the most significant number of students representing different faculties. Each course has (1) **Logs File** that contains entry time for e-learning events that were performed using specific e-learning components, a description of what the users are doing in each entry, how to access the system, and the IP address of the device. (2) **Users File** that contains information about the students and teachers in terms of student ID, teacher ID, and e-mails. (3) **Grades File** for the students regarding quizzes, assignments, first exam, second exam, mid-exam, final exam, and final grades. The first two files are obtained from the Center for E-Learning and Open Educational Resources, while the last one is obtained from the Admission and Registration Unit. After that, we have merged the students' names from the users' file and related grades from the grades file. Then, to keep the students' names and IDs' privacy, we have replaced them with new numbers. Finally, we have

Fig. 1 Workflow of Building E-LearningDJUST



analyzed the log file contents to determine each student's essential features with their grades.

The header row of log files contains (1) **time**: time of entry¹ to the e-learning system. (2) **User full name**: name of the e-learning user, whether admin user, teacher or student. (3) **Affected user**: the affected user when performing a particular activity in the system. (4) **Event context**: the course name. (5) **Component**: a group of features in the system including activities, resources, and different ways to track the progress of students (e.g., Quiz, Assignment, URL, etc.). (6) **Event name**: the event performed by the user in the system, like an attempt to start the quiz and submit a certain assignment. (7) **Description**: more details about each entry. (8) **Origin**: indicates the type of system access whether by computer, smartphone, or together. (9) **IP address**: address the devices used. On the other hand, the header row of the user file contains: (1) **ID Number**: unique number for users. (2) **Name**: name of the user. (3) **Email**: email address for each user. The grades files that are obtained from the Admission and Registration Unit contain the following header row: (1) **ID Number**: unique number for students. (2) **Name**: student's name. (3) **Semester Work Grades**: represents the semester work

grades during the semester (total grade before the final exam). (4) **Final Exam Grades**: represents the Final exam grades. (5) **Total out of 100 Grades**: represents the summation of semester work and final exam grades out of 100. (6) **Results**: means the results statutes included pass, fail, withdrawn, incomplete or absent.

The number of faculties is 13 at JUST, but our study excludes the graduate and research faculties. Therefore, the study concentrates on 11 faculties: Science and Arts, Pharmacy, Applied Medical Sciences, Agriculture, Medicine, Veterinary Medicine, Engineering, Nurse, Dentistry, Computer Information Technology, and Architecture and Design. To represent JUST students from these faculties, we have chosen four courses with large diverse students during the spring and summer semesters of 2019/2020 and fall semester of 2020/2021; we name them spring 2020, summer 2020, and fall 2021, respectively. The courses are CHEM101, CHEM262, CIS099, and PHY103. More details in the following sections.

3.1 Dataset description

The collected data are from courses of two faculties: Science and Arts, and Computer Information Technology (Computer & Info Tech) with different departments, as

¹ Entry: the record for each student in the e-learning system.

Table 1 Faculties and departments of active courses

Course id	Faculty	Department
CHEM101	Science and Arts	Chemistry
CHEM262	Science and Arts	Chemistry
CIS099	Computer and Information Technology	Computer Information System
PHY103	Science and Arts	Physics

Table 2 Number of students in active courses at three semesters

Course name	Course id	#Stu Spring	#Stu Summer	#Stu Fall
General chemistry	CHEM101	292	132	883
Biochemistry	CHEM262	612	1515	921
Computer skills	CIS099	876	1146	1172
General Physics	PHY103	600	633	464
Total Number of Students		2380	3426	3440

shown in Table 1. These courses are chosen since they are the standard university compulsory and elective courses that the students from 11 faculties were registered in.

Table 2 shows the courses name at each semester, course id, and the students’ distribution in each course at three semesters, respectively. The total number of students is 9,246 students distributed as follows: 2,380 students in spring 2020, 3,426 students in summer 2020, and 3,440 in fall 2021.

3.2 Dataset distribution

Figure 2 shows the number of students in each faculty mentioned above for the courses at spring, summer, and fall semesters.

The distribution of students in the spring semester is as follows: 313 in Agriculture, 309 in Applied Medical Sciences, 14 in Architecture and Design, 69 in Computer & Info Tech, 26 in Dentistry, 309 in Engineering, 114 in Medicine, 209 in Nurse, 615 in Pharmacy, 301 in Science and Arts, and 101 in Veterinary Medicine. While, the number of students in each faculty for the summer courses semester as follows: 467 in Agriculture, 836 in Applied Medical Sciences, 15 in Architecture and Design, 46 in Computer & Info Tech, 108 in Dentistry, 203 in Engineering, 130 in Medicine, 460 in Nurse, 722 in Pharmacy, 304 in Science and Arts, and 135 in Veterinary Medicine. Finally, the distribution of students in the Fall semester is as follows: 410 in Agriculture, 713 in Applied Medical Sciences, 24 in Architecture and Design, 90 in Computer & Info Tech, 11 in Dentistry, 848 in Engineering, 62 in Medicine, 362 in Nurse, 380 in Pharmacy, 504 in Science and Arts, and 36 in Veterinary Medicine.

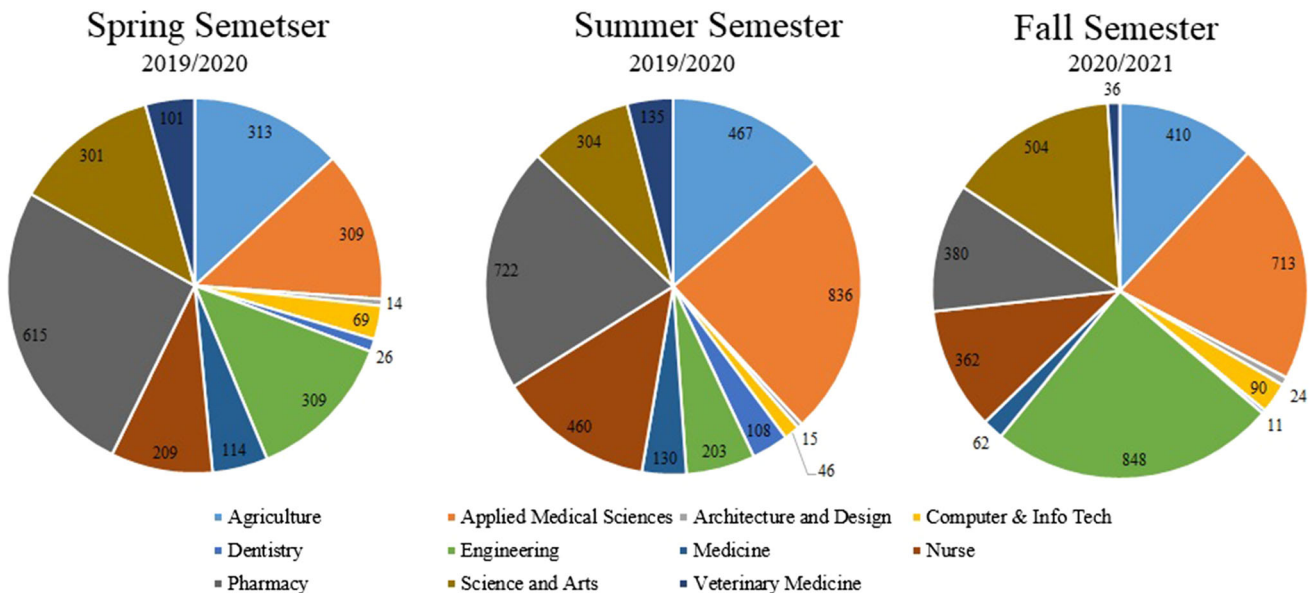


Fig. 2 Distribution of Students in each Faculty of three Semesters

Medicine, 362 in Nurse, 380 in Pharmacy, 504 in Science and Arts, 36 in Veterinary Medicine.

3.3 Dataset analyzing and filtering

The log file's main contents are entry time to the system, the components used, and the events attached to each user's entry. The components, such as a quiz, chat, or others, can be used by different users or the system itself. Also, each component has many events, such as submitting the quiz or opening the report performed by students, teachers, admin users, or the system. Table 3 shows the number of entries, components, and events that are related to students only.

After analyzing log files' content, the features were extracted. These features: the **Faculty** for each student, the **way used** to access the e-learning system either through their computer (web) or phone (ws), the number of **discussions** viewed for each of them, number of **submitted assignments**, the number of views of **activities and resources** for each course, number of **started, submitted, summary viewed, and viewed of each quiz**, number of

message sent between teachers and students or between students, number of views for their **grades or grades' reports**, number of **comments** that created by students of each assignment, the total number of events during the whole of the semester, and number of times that each student performed different activities per **week** during the semester. Finally, we concatenated student's features with their grades based on the users' files.

The current study examined the correlation between the events and the total grades (out of 100). Table 3 shows the five most important events (features) that have a high correlation score with total grades. In the Spring Semester Courses of 2020, the Quiz attempt started in CHEM262 has the highest impact on students' performances. For Summer courses, the Total number of events is essential to determine the student's study progress. Finally, for the Fall courses, the Quiz attempt viewed event in CHEM101 is vital for predicting students' performances.

Table 3 Number of entries, components, and events for all students

Course id	No. of Entries	No. of Components	No. of Events	Most 5 important events using RF feature selection
Spring semester courses				
CHEM101	15,979	8	15	Total events (0.31) The status of the submission.. (0.11) W13 (0.10) W17 (0.08) W15 (0.07)
CHEM262	79,877	7	9	Quiz attempt started (0.57) Quiz attempt viewed (0.14) W13 (0.07) W12 (0.07) Faculty (0.05)
CIS099	97,195	7	7	Quiz attempt submitted (0.42) Quiz attempt started (0.24) Quiz attempt viewed (0.14) Total Events (0.05) Course module viewed (0.04)
PHY103	79,697	9	15	Total Events (0.28) Week16 (0.17) Course module viewed (0.11) Faculty (0.07) Grade user report viewed (0.06)
Summer Semester Courses				
CHEM101	1986	7	8	Week7 (0.25) Clicked join meeting button (0.13) Week6 (0.09) Course module instance list viewed (0.08) Sessions viewed (0.07)
CHEM262	360,845	6	9	Course module viewed (0.33) Quiz attempt viewed (0.21) Week3 (0.19) Total Events (0.6) Faculty (0.5)
CIS099	159,311	8	10	Total Events (0.62) Quiz attempt viewed (0.15) Week5 (0.05) Chapter viewed (0.4) Quiz attempt submitted (0.3)
PHY103	113,447	7	8	Week5 (0.23) Course module viewed (0.13) Quiz attempt viewed (0.12) Week3 (0.10) Week4 (0.09)
Fall semester courses				
CHEM101	129,681	5	7	Quiz attempt viewed (0.52) Week10 (0.32) Week12 (0.03) Week13 (0.02) Week15 (0.02)
CHEM262	64,039	7	7	Course module viewed (0.26) Week8 (0.19) Faculty (0.12) Total Events (0.11) Week5 (0.06)
CIS099	196,370	9	10	Quiz attempt submitted (0.42) Total Events (0.15) Course module viewed (0.14) Week6 (0.13) Quiz attempt viewed (0.07)
PHY103	91,446	8	13	Faculty (0.31) Quiz attempt viewed (0.24) Week8 (0.09) Submission created (0.08) Week9 (0.05)

4 Experimental results

We have used regression models since the output variable (total grade out of 100) is a continuous real value. Using the E-LearningDJUST dataset, we have applied two machine learning models: Random Forest (RF), eXtreme gradient boosting (XGBoost), and one deep learning model, feed-forward neural network (FFNN).

We also used different regression metrics to evaluate the model performance and apply a comparison between them. These metrics are divided into two groups: (1) metrics for finding the errors between actual and prediction final grades and (2) metrics for finding the correlation between actual and prediction final grades.

1. *Regression error metrics* There are several metrics that find the errors between actual and prediction final grades: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Scatter index.
2. *Correlation metrics* There are several metrics that find the correlation between actual and prediction final grades: Pearson Correlation (Pear corr), Spearman Correlation (Spear corr), and R-square (R^2).

4.1 Random forest

Random forest[23] is a supervised ensemble learning method that relies on the decision tree. The class with the most votes becomes the model’s prediction in classification tasks. While in regression tasks, it computes the average prediction of all trees to get the model’s prediction. To build an RF model, it is necessary to adjust a model’s

hyperparameters. The number of trees is the number of estimators that are given as a parameter in the RF model. RF is not only prominent by its fast and efficient work in classification and regression tasks but is also able to arrange the importance of data features (see Table 4).

4.2 eXtreme gradient boosting (XGBoost)

XGBoost is an ensemble model developed to solve classification and regression problems based on a gradient boosting algorithm [24]. It contains several weak learners to generate a single strong learner. The weak learners are Gradient Boosting decision trees, in which each tree is performed individually, produces individual predictions, and then combines these predictions to form a final model’ prediction. It has a better control against overfitting by using more regularized model formalization in comparison to prior algorithms. The XGBoost also needs to be fine-tuned using hyperparameters (see Table 4).

4.3 Feed-forward neural network (FFNN)

FFNN is one of the machine learning methods that is based on the Artificial Neural Network. The structure of FFNN used in this study consists of three layers. The first layer is the input layer, which receives the features as input, and the last layer is the output layer, which produces the prediction result. In addition, there is one hidden layer for calculating and propagating weights. It is a feed-forward neural network with no backward connections between the neurons from the output layer to the hidden layer, and there is not any connection between the neurons in the same layer.

Table 4 Parameters and values of baseline models

Baseline model	Parameter	Value	
RF	random_state	42	
	max_features	auto , log2, sqrt	
	n_estimators	50 , 100, 500	
	max_depth	2 , 5, 10	
XGB	random_state	42	
	colsample_bytree	1 , 0.9, 0.7	
	eta	0.3 , 0.1, 0.4, 0.5	
	max_depth	2 , 4, 6, 8	
FFNN	subsample	0.9 , 1.0	
	Input Layer	Dense Layer	units = 64 , 32, 16 activation = ReLU
	–	Drop out	0.2 , 0.1
	Hidden Layer	Dense Layer	units = 16 , 8 activation = ReLU
	Output Layer	Dense Layer	units = 1
	epochs		500 , 50, 100
	batch size		20 , 32
validation_split		0.2 , 0.3	

Table 5 Results of machine learning models for CHEM101, CHEM262, CIS099, and PHY103 courses

Course ID	Model	RMSE	MAE	MAPE	R2	Scatter Index	Pear corr	Spear corr
Spring semester 2019/2020								
CHEM101	RF	11.02	8.60	14.30	0.03	17.97	0.23	0.14
	XGB	11.14	8.47	14.07	0.01	18.17	0.20	0.20
	FFNN	14.85	11.51	19.37	− 0.76	24.22	0.11	0.06
CHEM262	RF	8.08	6.46	8.60	0.18	10.47	0.44	0.33
	XGB	7.98	6.35	8.36	0.20	10.35	0.45	0.27
	FFNN	10.31	7.99	9.95	− 0.34	13.37	0.32	0.16
CIS099	RF	10.2	8.0	13.36	0.18	15.55	0.43	0.44
	XGB	9.77	7.7	12.78	0.24	14.9	0.49	0.47
	FFNN	11.63	9.88	14.81	− 0.07	17.74	0.67	0.67
PHY103	RF	15.13	11.30	0	0.19	23.43	0.45	0.39
	XGB	14.73	10.96	0	0.23	22.82	0.49	0.42
	FFNN	17.18	13.03	inf	− 0.05	26.6	0.34	0.33
Summer Semester 2019/2020								
CHEM101	RF	8.75	6.92	8.51	− 0.24	10.34	− 0.10	− 0.04
	XGB	8.64	6.59	8.03	− 0.21	10.22	0	0.06
	FFNN	13.02	10.68	12.78	− 1.74	15.39	0.1	0.07
CHEM262	RF	7.09	5.49	8.30	0.02	10.26	0.17	0.18
	XGB	7.25	5.60	8.38	− 0.02	10.49	0.14	0.13
	FFNN	9.64	7.76	11.01	− 0.81	13.95	0.2	0.17
CIS099	RF	8.91	7.08	12.16	0.12	13.85	0.38	0.33
	XGB	8.55	6.76	11.48	0.19	13.31	0.45	0.37
	FFNN	10.67	8.51	13.59	− 0.26	16.6	0.26	0.23
PHY103	RF	11.13	8.91	12.90	− 0.01	14.74	0.05	0.06
	XGB	11.13	8.79	12.66	− 0.01	14.73	0.06	0.05
	FFNN	10.26	8.46	11.85	− 0.05	13.58	0.21	0.24
Fall Semester 2020/2021								
CHEM101	RF	12.93	10.44	17.98	0.06	19.86	0.26	0.28
	XGB	12.96	10.54	18.41	0.05	19.91	0.30	0.32
	FFNN	13.86	11.14	19.15	− 0.08	21.29	0.2	0.15
CHEM262	RF	8.07	6.26	10.05	0.06	11.90	0.30	0.20
	XGB	7.74	6.07	9.62	0.14	11.41	0.40	0.29
	FFNN	10.92	9.01	13.13	− 0.72	16.09	0.28	0.16
CIS099	RF	9.13	7.24	11.80	0.14	14.33	0.38	0.36
	XGB	9.22	7.36	11.86	0.13	14.47	0.36	0.34
	FFNN	11.43	9.14	13.77	− 0.35	17.95	0.22	0.16
PHY103	RF	12.42	9.87	14.44	0.16	16.45	0.42	0.44
	XGB	12.42	9.64	13.99	0.16	16.45	0.40	0.41
	FFNN	14.99	11.88	16.55	− 0.23	19.86	0.25	0.29

More details about the finetuning hyperparameters are shown in Table 4.

4.4 Results of models

Table 5 shows the experimental results for the three models after determining the parameters with the corresponding values, as shown in Table 4. It is clear from the table that RF and XGB have better results than FFNN in the spring

and fall semesters. While in the summer semester, the FFNN shows a small improvement over RF and XG in PHY103.

As a results explanation, based on the correlation between events and the total score (out of 100), we can observe that the CIS099 and CHEM262 subjects have the highest correlations. Thus, both have high predictive scores compared to the CHEM101 and PHY103 subjects. This indicates that the more the student uses the e-learning

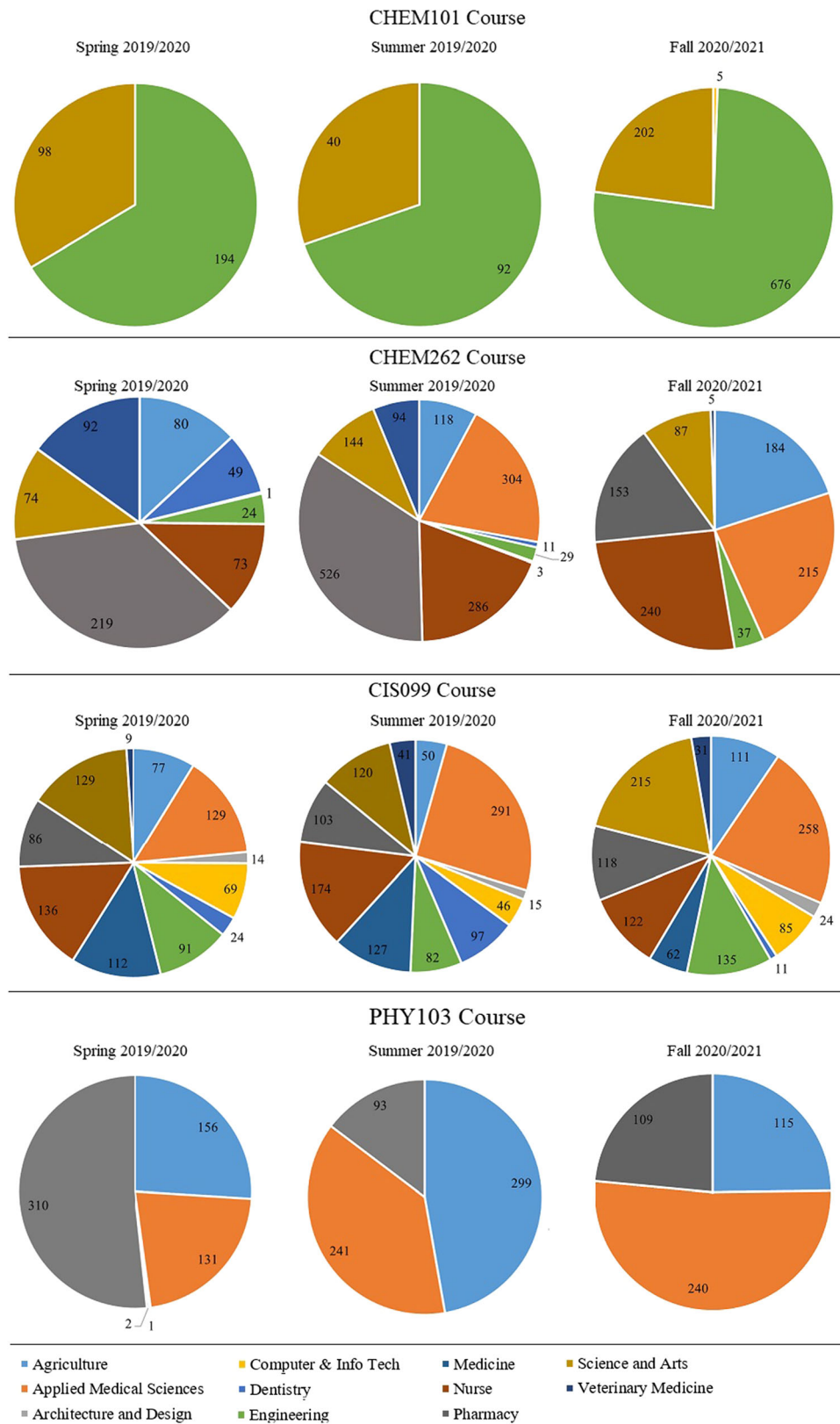


Fig. 3 Distribution of Students in each Faculty of each course

system, the more he obtains a high grade. Also, knowing that CHEM101 and PHY103 have lower student diversity than CHEM262 and CIS099 (see Fig. 3), this may indicate that specialist subjects, such as CHEM101 and PHY103 are more challenging to predict than subjects whose students are from various undergraduate majors.

We have also applied RF with feature selection using features selection and compare the results with 'auto' parameters. The results in Table 6 suggests that building a model from the most important features alone results in a more effective model. This could be because the other features are redundant and need to be removed.

5 Coronavirus implication

The sudden emergence of COVID– 19 has forced educational institutions to shift education to online platforms for the safety of students and teachers. It is satisfying to know that students in this educational institution were able to access educational platforms, communicate with their teachers and follow up on the materials. This was evident from the number of times they accessed the e-learning platform. But it must be noted that exams without supervision have greatly affected the process of evaluating students. There is no doubt that cheating methods increase in the absence of supervision over exams [25–27].

To study the impact of the Coronavirus pandemic on students' performances and how the students benefited from the online learning, we obtained grade datasets out of 100 for the same selected courses before the appearance of Coronavirus semesters (Fall 2018/2019, Spring 2018/2019, Summer 2018/2019, and Fall 2019/2020). In addition to the semesters during the pandemic (Spring 2020, Semester 2020, and Fall 2021). This new dataset was also collected from the Admission and Registration Unit at JUST. Several statistical experiments were performed based on the number of students over seven semesters for three academic years: four semesters before the Coronavirus pandemic and

three during it. The statistical experiments carried out in this study are divided into three groups: (1) the changes in the number of students, (2) the rate of passes and fails, and 3) the number of students in each passing level (90–100, 80–89, 70–79, 60–69, 50–60, and <50).

Regarding the number of students in each class, Table 7 shows all the numbers for all semesters. For CHEM262 and PHY103 courses, the number of students before and during the pandemic was almost the same. However, the number of students in the last fall semester of CHEM101 was noticeably less than the regular semesters. On the other hand, the number of students in the CIS099 course increased during the pandemic. This course was recently renamed CIS099 (formerly CIS100); Thus, in previous semesters, many students registered for CIS100. This is the explanation for the massive increase in student numbers during the pandemic.

1. CHEM101 course

Figure 4 shows the percentage of the pass and fail for CHEM101 in all semesters. The figure indicates that the students' performances during the Coronavirus pandemic had increased compared with their previous performances. Therefore, we can conclude that the Coronavirus affects students' performance by increasing the rate of passing students in the CHEM101

Table 7 Number of students in each semester between 2018 and 2021 for the four courses

Semester	CHEM101	CHEM262	CIS099	PHY103
Fall 2018/2019	1002	933	379	501
Spring 2018/2019	328	820	469	485
Summer 2018/2019	121	546	430	433
Fall 2019/2020	1381	757	622	393
Spring 2019/2020	292	612	876	600
Summer 2019/2020	132	1515	1146	633
Fall 2020/2021	883	921	1172	464

Table 6 RMSE results using Random Forest with all features versus Random Forest with feature selection of best features with threshold importance ≥ 0.02

RMSE for Random Forest	Spring 2020		Summer 2020		Fall 2021	
	Using all Features	Feature Selection	Using all Features	Feature Selection	Using all Features	Feature Selection
CHEM101	11.02	10.94	8.75	8.73	12.93	12.91
CHEM262	8.08	8.07	7.09	7.02	8.07	8.05
CIS099	10.2	10.18	8.91	8.91	9.13	9.13
PHY103	15.13	14.93	11.13	11.13	12.42	12.38

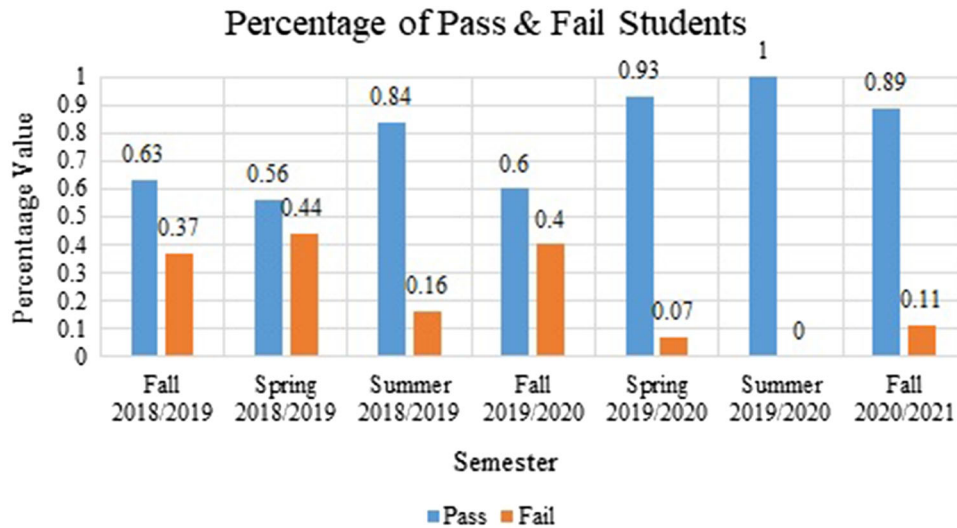


Fig. 4 Percentage of Pass & Fail Students in CHEM101 course

course in the Spring, Summer semesters of 2020, and Fall 2021.

We have also divided the pass group into five levels: 90-100, 80-89, 70-79, 60-69, and 50-59. Then, we computed the percentage of students in each level, as shown in Fig. 5. It can be seen from the figure that students’ grades distributions during the pandemic have different distributions than before the pandemic. Knowing that the number of students in the Fall is larger than the number of students in the Spring and Summer semesters, concentrating on the Fall, the distribution of the grades in Fall 2021 shows a normal distribution. Knowing that the exams were online,

noting a normal distribution for the grades indicates that online teaching is progressing toward evaluating students and preventing cheating.

2. CHEM262 course

Figure 6 shows the percentage of passing and fail in CHEM262 for all semesters. Although the number of failed students is the lowest during the online teaching, the grade distribution in Fig. 7 shows that most of the grades are between 60 and 80.

3. CIS099 course

Figure 8 shows the pass and failure percentages in CIS099. While Fig. 9 shows the grades distributions. It is

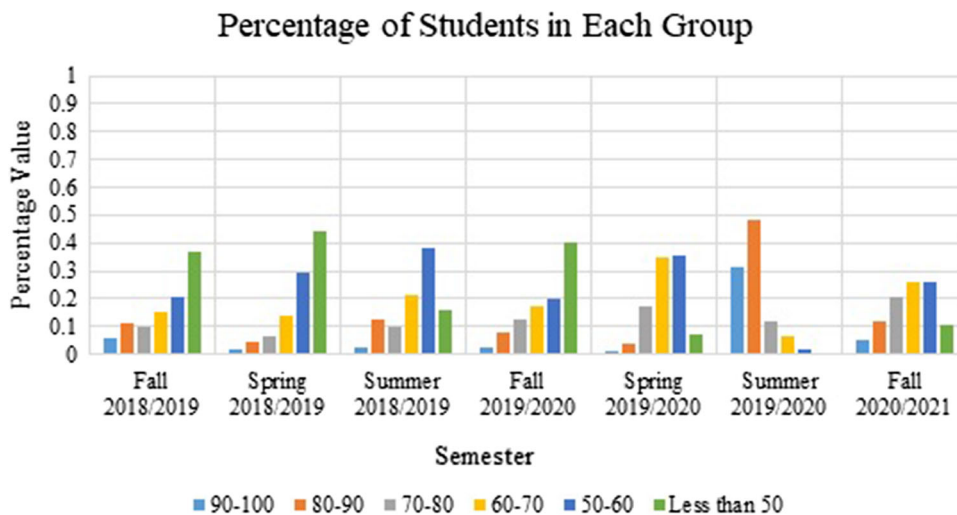


Fig. 5 Percentage of Pass Students Levels in CHEM101 course

Fig. 6 Percentage of Pass & Fail Students in CHEM262 course

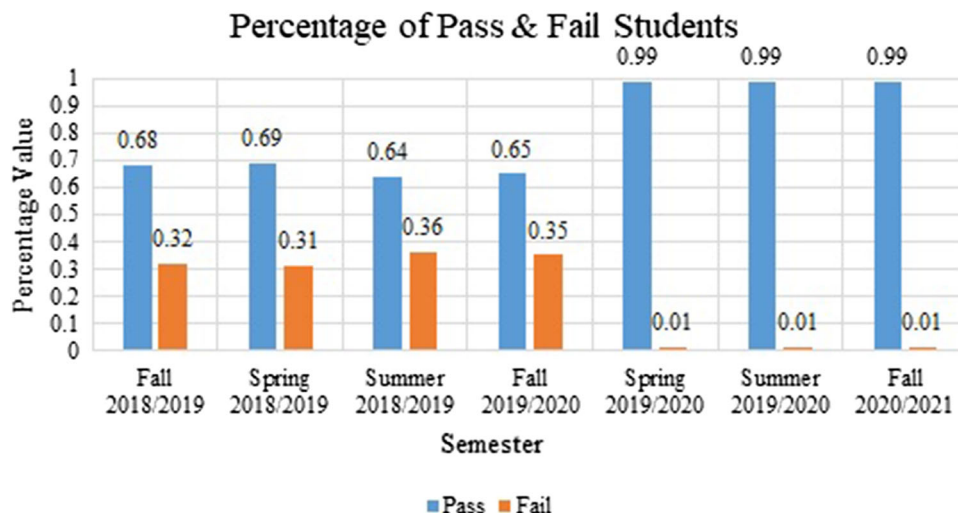
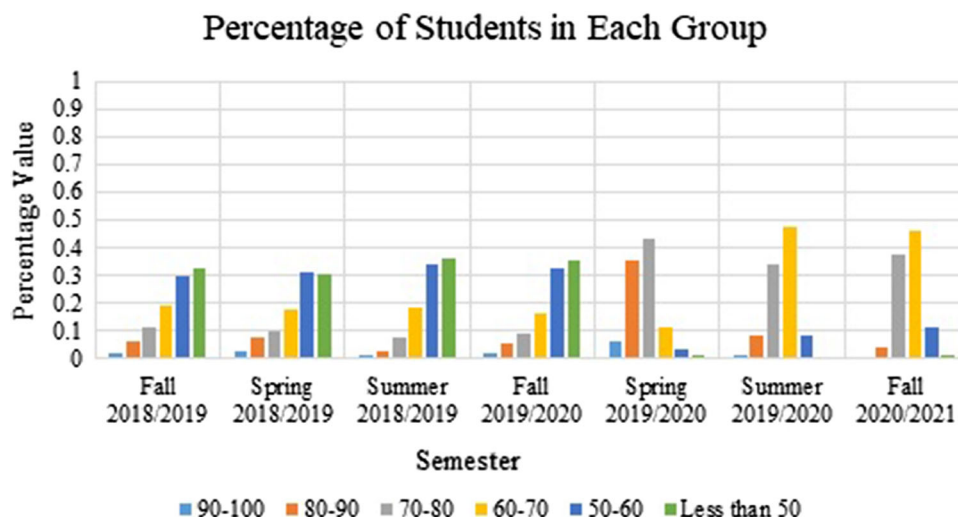


Fig. 7 Percentage of Pass Students Levels in CHEM62 course



worth mentioning that students can skip this course if they pass the Computer Skills exam once they are admitted into the school. If the student doesn't pass the exam, then the student should take this course. This course was recently renamed CIS099 (formerly CIS100); Thus, in previous semesters, many students registered for CIS100. This is the explanation for the massive increase in student numbers during the pandemic.

Since we didn't collect the data for CIS100, we could not compare the pass and fail rates in CIS099 in different semesters. However, the normal distribution of the grades during the Coronavirus pandemic is remarkable.

4. PHY103 course

The students were divided into two groups based on the final grades in the PHY103 course: Pass or Fail. First, we computed the percentage of students in each group, as shown in Fig. 10. After dividing the students into the pass and fail groups, we divided the pass group into five levels: 90–100, 80–89, 70–79, 60–69, and 50–59. Then, we computed the percentage of students in each level, as shown in Fig. 11. It has been noticed that the rate of passing increased during the pandemic.

This study found that the percentage of students with high scores above 80 was similar to before the pandemic. This may be an indication that online learning can distinguish high achieving students. On the other hand, the percentage of students who scored 50–75 during the

Fig. 8 Percentage of Pass & Fail Students in CIS099 course

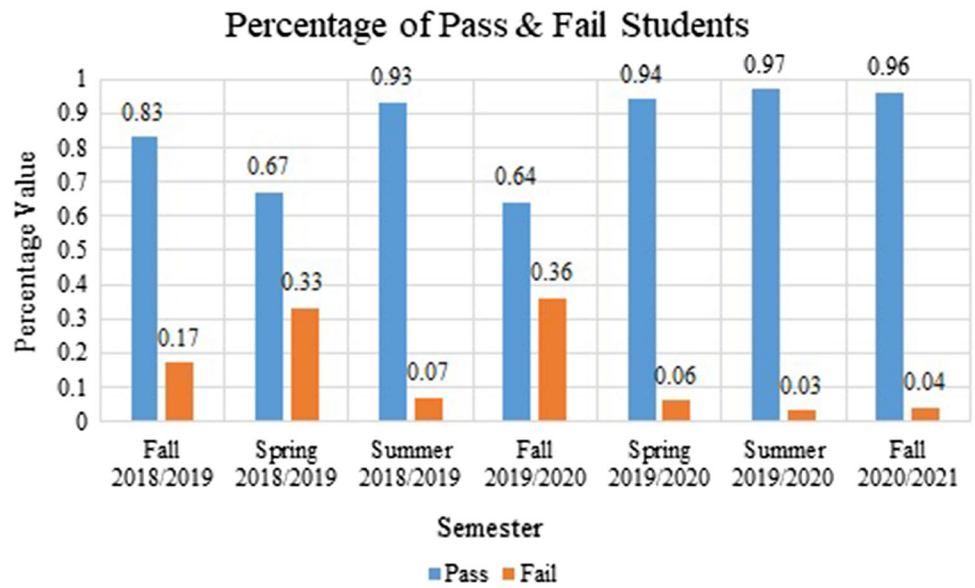


Fig. 9 Percentage of Pass Students Levels in CIS099 course

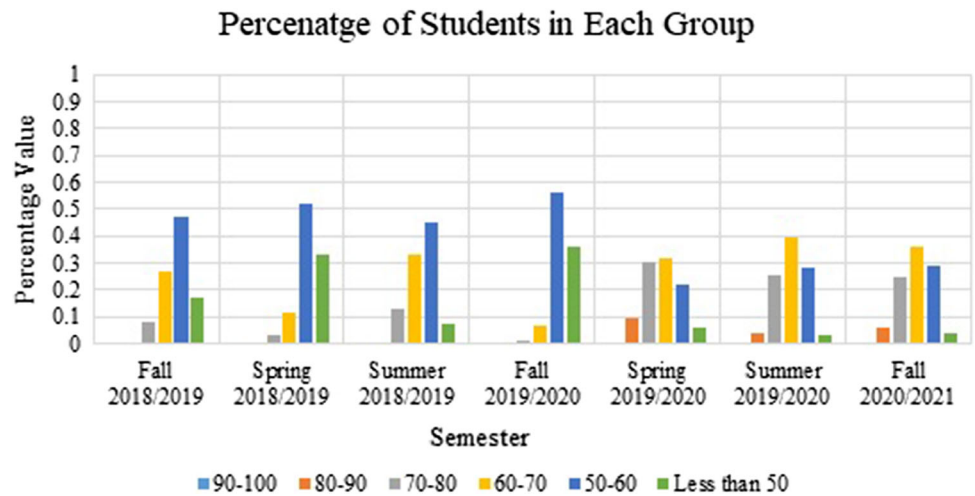
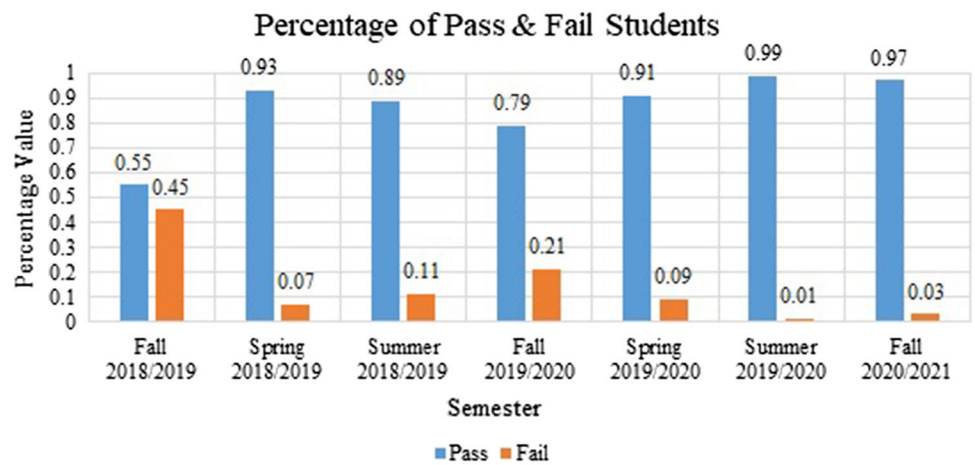


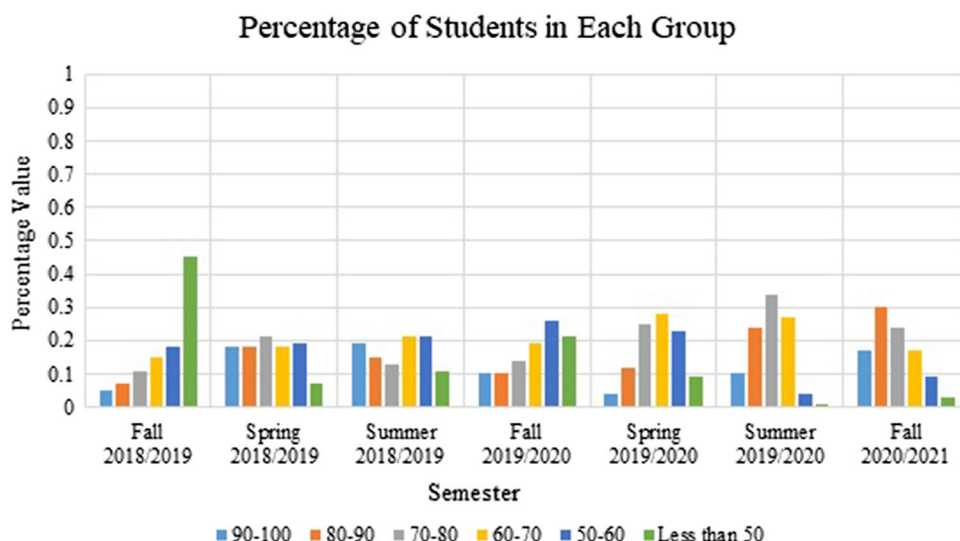
Fig. 10 Percentage of Pass & Fail Students in PYH103 course



pandemic was higher than before. Here, it is worth noting that replacing traditional exams with online assessment

tools is still a new experience for many educational institutions in Jordan due to the pandemic. Therefore, student

Fig. 11 Percentage of Pass Students Levels in PYH103 course



performance evaluation is still under study. It is undeniable that the absence of invigilators during the exams may create an excuse for some students to practice different methods to achieve success. Unfortunately, it will be difficult to identify these methods. We also believe that some students may perform better when taking their exams in a more relaxed home atmosphere than college classes under the pressure of the invigilators' presence. Admittedly, one of the disadvantages and challenges of online learning is the student assessment process. As a new procedure for the new academic year 2021/2022, universities in Jordan offer three types of classes; online, hybrid, and in-person, with on-campus assessment exams taking place for most subjects.

6 Conclusion

This research paper introduces and examines the E-LearningDJUST dataset representing students' study progress at Jordan University of Science and Technology (JUST) for three semesters (spring, summer of 2019/2020, and fall of 2020/2021). The dataset depicts a sample of the university's students. This dataset is considered the first collected dataset that reflects the student's study progress within a Jordanian institute using an e-learning system database to the best of our knowledge. The log and user files are obtained from the Center for e-learning and Open Educational Resources. In addition, the grades of students are collected from the Admission and Registration Unit. The dataset contains 9,246 students distributed over three semesters of the two academic years 2019/2020 and 2020/2021: 2,380 in the spring semester, 3,426 in the summer semester, and 3,440 in the fall semester. These students are registered in 11 faculties.

We analyzed and filtered the dataset in which we extracted features for students only to predict their performances. The featured extracted, and the total final grades were found to correlate, indicating that the e-learning usage impacted the students' performances. Furthermore, the results showed that the quiz with related events had the most impact on the performances. We also applied three models on the E-LearningDJUST dataset: two machine learning models (RF and XGB) and one deep learning model (FFNN). The best results for these models came from RF and XGB models. Also, the result of RF shows more improvements when it was applied on less number of features using RF selected feature. For example, in CHEM101 (Spring 2020), applying RF on all features resulted in 11.02 RMSE. At the same time, applying RF with the essential features resulted in 10.94 as RMSE.

Finally, we conducted a statistical experiment to see if the Coronavirus impacted the students' performance by counting the number of students in different groups of grades before and during the Coronavirus pandemic. We also examined the total number of students in each course at seven semesters. In addition, the study showed the number of passes and failed students in each course at seven semesters for the academic year (2018–2021). Finally, we illustrated the number of students in pass levels (90–100, 80–89, 70–79, 60–69, 50–60) and one fail level (grade less than 50). The results showed that that the rate of passing increased during the pandemic. However, the grades distribution is considered normal distribution for most of the cases during the pandemic. Although online teaching during the pandemic has shown remarkable success, we firmly believe that giving exams inside campus would prevent any cheating that may cause the change of success and fail rates. The research proved that most students benefited from online learning, but the

recommendations confirm, as in many studies [25–27], that there should be different ways to evaluate students, such as the student's attendance at the exam venue or monitoring the activities of the device he uses for the exam and making the students use cameras during the exam.

Acknowledgments We gratefully acknowledge the Deanship of Research at the Jordan University of Science and Technology (JUST) for supporting this work via Grants #20200064 and #20210039. We also recognize the efforts of the Center for e-learning and Open Educational Resources and the Admission and Registration Unit at JUST for providing us with the dataset.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

- AlHamad AQM (2020) Acceptance of e-learning among university students in UAE: a practical study. *Int J Electr Comput Eng* (2088–8708) 10(4):3660–3671
- Bakhouyi A, Dehbi R, Talea M, Hajoui O (2017) 16th international conference on information technology based higher education and training (ITHET), IEEE, pp 1–8
- Ennouamani S, Mahani Z (2017) 2017 eighth international conference on intelligent computing and information systems (ICICIS), IEEE, pp 342–347
- Hussain M, Zhu W, Zhang W, Abidi SMR, Ali S (2019) Using machine learning to predict student difficulties from learning session data. *Artif Intel Rev* 52(1):381–407
- Wang M (2018) *E-learning in the workplace*. Springer, Berlin, pp 41–53
- Sidey-Gibbons JA, Sidey-Gibbons CJ (2019) Machine learning in medicine: a practical introduction. *BMC Med Res Methodol* 19(1):1–18
- Abedalla A, Abdullah M, Al-Ayyoub M, Benkhelifa E (2021) Chest x-ray pneumothorax segmentation using u-net with efficientnet and resnet architectures. *PeerJ Comput Sci* 7:e607
- Ali A, Zhu Y, Zakarya M (2021) A data aggregation based approach to exploit dynamic spatio-temporal correlations for citywide crowd flows prediction in fog computing. *Multimedia Tools Appl*, pp 1–33
- Ali A, Zhu Y, Chen Q, Yu J, Cai H (2019) 2019 IEEE 25th international conference on parallel and distributed systems (ICPADS), IEEE, pp 125–132
- Ali A, Zhu Y, Zakarya M (2021) Exploiting dynamic spatio-temporal correlations for citywide traffic flow prediction using attention based neural networks. *Inf Sci* 577:852–870
- Abedalla A, Fadel A, Tuffaha I, Al-Omari H, Omari M, Abdullah M, Al-Ayyoub M (2019) 2019 sixth international conference on social networks analysis, management and security (SNAMS), IEEE, pp 274–278
- Villegas-Ch W, Román-Cañizares M, Palacios-Pacheco X (2020) Improvement of an online education model with the integration of machine learning and data analysis in an lms. *Appl Sci* 10(15):5371
- Zhai X, Yin Y, Pellegrino JW, Haudek KC, Shi L (2020) Applying machine learning in science assessment: a systematic review. *Stud Sci Edu* 56(1):111–151
- Sana B, Siddiqui IF, Arain QA (2019) Analyzing students' academic performance through educational data mining
- Iatrellis O, Savvas IK, Fitsilis P, Gerogiannis VC (2021) A two-phase machine learning approach for predicting student outcomes. *Educ Inf Technol* 26(1):69–88. <https://doi.org/10.1007/s10639-020-10260-x>
- Aggarwal D, Mittal S, Bali V (2021) Significance of non-academic parameters for predicting student performance using ensemble learning techniques. *Int J Syst Dyn Appl* 10(3):38–49. <https://doi.org/10.4018/ijdsda.2021070103>
- González MR, de Puerto Paule Ruíz M, Ortin F (2021) Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Comput Educ* 163:104–108. <https://doi.org/10.1016/j.compedu.2020.104108>
- Conijn R, Snijders C, Kleingeld A, Matzat U (2017) Predicting student performance from LMS data: a comparison of 17 blended courses using moodle LMS. *IEEE Trans Learn Technol* 10(1):17–29. <https://doi.org/10.1109/TLT.2016.2616312>
- Hamoud AK, Hashim AS, Awadh WA (2018) Predicting student performance in higher education institutions using decision tree analysis. *Int J Interact Multim Artif Intell* 5(2):26–31. <https://doi.org/10.9781/ijimai.2018.02.004>
- Hashim AS, Awadh WA, Hamoud AK (2020) IOP conference series: materials science and engineering, vol 928, IOP Publishing, p 032019
- Abu-Naser SS, Zaqout IS, Abu Ghosh M, Atallah RR, Alajrami E (2015) Predicting student performance using artificial neural network: in the faculty of engineering and information technology. *Int J Hybrid Inf Technol* 8(2):221–228
- Zeineddine H, Braendle U, Farah A (2021) Enhancing prediction of student success: automated machine learning approach. *Comput Electr Eng* 89:106–903. <https://doi.org/10.1016/j.compeleceng.2020.106903>
- Ho TK (1995) Proceedings of 3rd international conference on document analysis and recognition, vol 1, IEEE, pp 278–282
- Chen T, Guestrin C (2016) Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785–794
- Ali L, Dmour N (2021) The shift to online assessment due to covid-19: an empirical study of university students, behaviour and performance, in the region of uae. *Int J Inf Edu Technol* 11(5):220–228
- Hill G, Mason J, Dunn A (2021) Contract cheating: an increasing challenge for global academic community arising from covid-19. *Res Practice Technol Enhanced Learn* 16(1):1–20
- Bilen E, Matros A (2021) Online cheating amid covid-19. *J Econ Behav Organ* 182:196–211

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.