Computers and Education: Artificial Intelligence 3 (2022) 100077

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Appraisal of high-stake examinations during SARS-CoV-2 emergency with responsible and transparent AI: Evidence of fair and detrimental assessment

MD. Rayhan [a], MD. Golam Rabiul Alam [a], M. Ali Akber Dewan [c,*], M. Helal Uddin Ahmed [b]

[a] Department of Computer Science and Engineering, Brac University, 66 Mohakhali, Dhaka, 1212, Dhaka, Bangladesh
[b] Department of Management Information Systems, University of Dhaka, Dhaka, 1000, Dhaka, Bangladesh
[c] School of Computing and Information Systems, Faculty of Science and Technology, Athabasca University, Edmonton, AB, T5J 3S8, Canada

ARTICLE INFO

Keywords:
Evaluation methodologies
Automated assessment
Covid-19 education response
Explainable AI
High-stakes examination
AI in Education

ABSTRACT

In situations like the coronavirus pandemic, colleges and universities are forced to limit their offline and regular academic activities. Extended postponement of high-stakes exams due to health risk hereby reduces productivity and progress in later years. Several countries decided to organize the exams online. Since many other countries with large education boards had an inadequate infrastructure and insufficient resources during the emergency, education policy experts considered a solution to simultaneously protect public health and fully resume high-stakes exams -by canceling offline exam and introducing a uniform assessment process to be followed across the states and education boards. This research proposes a novel system using an AI model to accomplish the complex task of evaluating all students across education boards with maximum level of fairness and analyzes the ability to fairly appraise exam grades in the context of high-stakes examinations during SARS-CoV-2 emergency. Basically, a logistic regression classifier on top of a deep neural network is used to output predictions that are as fair as possible for all learners. The predictions of the proposed grade-awarding system are explained by the SHAP (SHapley Additive exPlanations) framework. SHAP allowed to identify the features of the students' portfolios that contributed most to the predicted grades. In the setting of an empirical analysis in one of the largest education systems in the Global South, 81.85% of learners were assigned fair scores while 3.12% of the scores were significantly smaller than the actual grades, which would have had a detrimental effect if it had been applied for real. Furthermore, SHAP allows policy-makers to debug the predictive model by identifying and measuring the importance of the factors involved in the model's final decision and removing those features that should not play a role in the model's "reasoning" process.

1. Introduction

The global health crisis resulting from the 2020 coronavirus outbreak caused the biggest worldwide education disruption in history. Extended postponement of high-stakes exams due to health risk prevented millions of examinees to proceed further in their career. Moreover, the prolonged suspension of exams can have a huge financial impact. These operational barriers may reduce university admissions and course enrollments (Dhanalakshmi et al., 2021). The virus outbreak has led governments around the globe to temporarily shut down kindergartens, schools, colleges, universities, and other learning institutions nationwide. The majority of the countries decided to continue the massive closure of schools[1] as the number of Covid-19 cases exponentially increased.[2] Thus, The lockdown and social distancing measures immediately had an enormous impact on education. The suspension of the education system was maintained during the development of the vaccines, which was expected to take many years (Niko Kommenda, 2020). The large education boards in the Global South were suspended till the midyear of 2021. Global and National policy-makers of education bodies were trying their best to tackle the unprecedented emergency in education. For instance, to ensure learning continuity, many countries transferred campus learning to online learning.

* Corresponding author. Faculty of Science and Technology, School of Computing and Information Systems, Athabasca University, Athabasca, AB, T9S 3A3, Canada.
E-mail addresses: md.rayhan@g.bracu.ac.bd (MD. Rayhan), rabiul.alam@bracu.ac.bd (MD.G.R. Alam), adewan@athabascau.ca (M.A.A. Dewan), helal@du.ac.bd (M.H.U. Ahmed).

[1] Global monitoring of school closures caused by Covid-19.
[2] WHO Health Emergency Dashboard.

https://doi.org/10.1016/j.caeai.2022.100077
Received 13 September 2021; Received in revised form 30 April 2022; Accepted 1 May 2022
Available online 20 May 2022

However, education bodies remained puzzled over how to handle the scheduled assessments and exams, especially the high-stakes exams such as school leaving, gateways for the job, and other major public exams handled by the central education board. The majority of high school graduates typically join the military or enroll in medical or engineering postsecondary education. Halting the graduation process will impact the future career of these students as well as the economy of the affected countries. Interruption of high-stakes exams delays student qualification and graduation, which in its turn delays entering higher education or the job market. Hence, throughout 2020 and 2021, handling high-stakes exams was among the top priorities of all policy-makers' agenda (UNESCO's COVID-19 Education Response, 2020). Lower and middle-income countries are more vulnerable to time and resource constraints in handling nationwide high-stakes exams (Davidson & Katopodis, 2020).

Various measures have been considered to cope with the emergency including cancellation, postponement, derogation, on-screen test, paper-based examinations with physical distancing, remote assessment, and using alternative approaches for validation and certification. Each of the solutions has its drawbacks in terms of fairness and evaluation quality; therefore, the set of solutions is not equally applicable to all education systems. If one set of solutions was feasible then the country's policy response would not be diverse.[3] At the beginning of the pandemic and school closure, education bodies around the world planned to arrange the high-stakes exams on the reopening of schools. As 2020 was reaching its end, with millions of candidates' progression already hampered, countries rapidly opted for new strategies regarding the assessment of high-stakes exams. During the absence of formal assessment, many country governments introduced and approved alternative approaches to high-stakes exams.

After a long period of waiting and lengthy consideration over accountability and health risk, one of the E9 countries,[4] Bangladesh, decided to assign the arithmetic mean of the previous two exams to the final exam of its graduating college students. Around the same time, the Indian Central Board of Secondary Education (CBSE) promulgated a 40:30:30 formula for evaluation of Class XII graduation exam (The Indian Express, 2021). The 40:30:30 is the percentage to be considered from class XII, XI and X result. Bangladesh and India declared that they had no other option but to consider an alternative instead of the offline college leaving exam to protect the safety, health, and social-emotional well-being of students and educational personnel as well as to alleviate the logistic and financial burdens associated with organizing and conducting exams. Adapting an evaluation system because of war or a major conflict is referred to as **latent** assessment strategy (Clarke, 2011). The latent assessment model for the Bangladesh Higher Secondary Certificate exam is a mathematical model that generates results looking at previous two achievements in the secondary stage. The purpose of an alternate appraisal model for high-stakes exam grades as an emergency measure is to maximize fairness and equity. The higher education board of Bangladesh applied a simple arithmetic function based on two parameters: performance in junior and senior secondary level. A mathematical model's abstraction about the real-world system is expressed in its conceptual model (Levins, 1966). Simple models with a small number of parameters are typically unaware of the probability distribution. Evaluating high-stakes exam candidates with such simpler models has a higher probability of distorting the true distribution as well as the fairness of evaluation. The country's experts believed-there was apparently no solution that would favor everybody amid too many uncertainties (Hashan, 2020). Several times before the All India Senior School Certificate Examination result publication, the high courts of India ordered CBSE to modify the proposed formula to ensure equity (News18, 2021b). The consequence of publishing high-stakes grades with arithmetic formula is quite upsetting when reportedly thousands of candidates challenged their awarded grade and extreme self-injurious attempt were even taken by some young adults (News18, 2021a). Not only does the detriment effect create unexpected consequences but the unexpectedly higher number of students with an awarded grade above 90% will make the university admission an uphill task. The arithmetic formula for high-stakes exam assessment made a large number of gainers and losers. This article identifies the research gap since there is no globally conceded model as an alternative to high-stakes exams during crises and, hereby proposes a robust framework that is capable of appraising grades of all types of candidates based on the student portfolio.

A student portfolio is a set of educational data including academic accomplishment, identification, awards, obtained marks, honors, certifications, etc. compiled in any form, presented publicly or privately. Recent literature shows that AI and educational data mining are research fields still in their infancy and their successful application in educational institutions has to be demonstrated more extensively (Saddiqa et al., 2021)(Lemay et al., 2021). The previous decade has witnessed a remarkable growth in research on computer-based evaluation based on a student portfolio (Chen, Zou, et al., 2020). However, little effort has been attempted to incorporate deep learning technology into formal assessment systems (Chen, Xie, et al., 2020). This research established an inclusive and equitable machine learning model beyond the idea of uniform assessment formula. The research question below drives the efforts of this study.

Is an AI-driven assessment system an effective pragmatic solution during the absence of formal high-stakes assessment activity?

This research problem became obvious to the author when the Global South nations were in quest of an inclusive assessment formula to assess high-stakes exams (News18, 2021c). While the method applies to any education system, this study exemplifies the method by applying it to one of the E−9 member country's higher school certificate assessment. Moreover, the model is applicable as a smart learning analytic tool (Chen & Li, 2021)(Solano-Flores et al., 1999) to support precision education (Yang et al., 2021, Table 4) which will promote fairer student assessment (Friedler et al., 2008). A systematic literature review discovered that early student performance prediction can help universities to provide timely actions, like planning for appropriate training to improve students' success rate (Alyahyan & Düştegör, 2020). The centralized high-stakes examinations are summative evaluations that are generally carried out at the end of the learning process. One disadvantage of this type of assessment is that the learners discover their true performance when it is too late. A reliably predicted grade sheet would enhance a student's confidence as well as suggest emphasizing particular subjects. Therefore, the prediction of academic performance in higher education provides several benefits to teachers, students, policy-makers, and institutions.

The proposed system is tested in the setting of an empirical study with the central board's Class 12 examination, which aims at determining the model's ability to fairly evaluate high-stakes exam grades, while optimizing the gainer-to-loser ratio. An example of a hypothetical dedicated computer to predict exam grades is illustrated in Fig. 1 works like a vending machine that outputs a grade per subject when some student portfolio is input. Later in Section3.1, the Turing test will be applied to the machine.

The current research on autonomous assessment systems is discussed in 2.1 followed by a discussion on the pandemic's global disruption of formal education and assessment in 2.2. The E−9 member country's secondary education system, in which this study's evaluation model is tested, is then shortly described in 2.3 followed by sections about emerging assessment policies during the pandemic 2.4 and AI systems in higher education 2.5. Building blocks or components of the proposed

[3] country's policy response and reopening plans tracker by Center for Global Development.

[4] Group E−9 countries (Bangladesh, Brazil, China, Egypt, India, Indonesia, Mexico, Nigeria, and Pakistan) represents over half of the world's population and some of the largest education systems in the world.

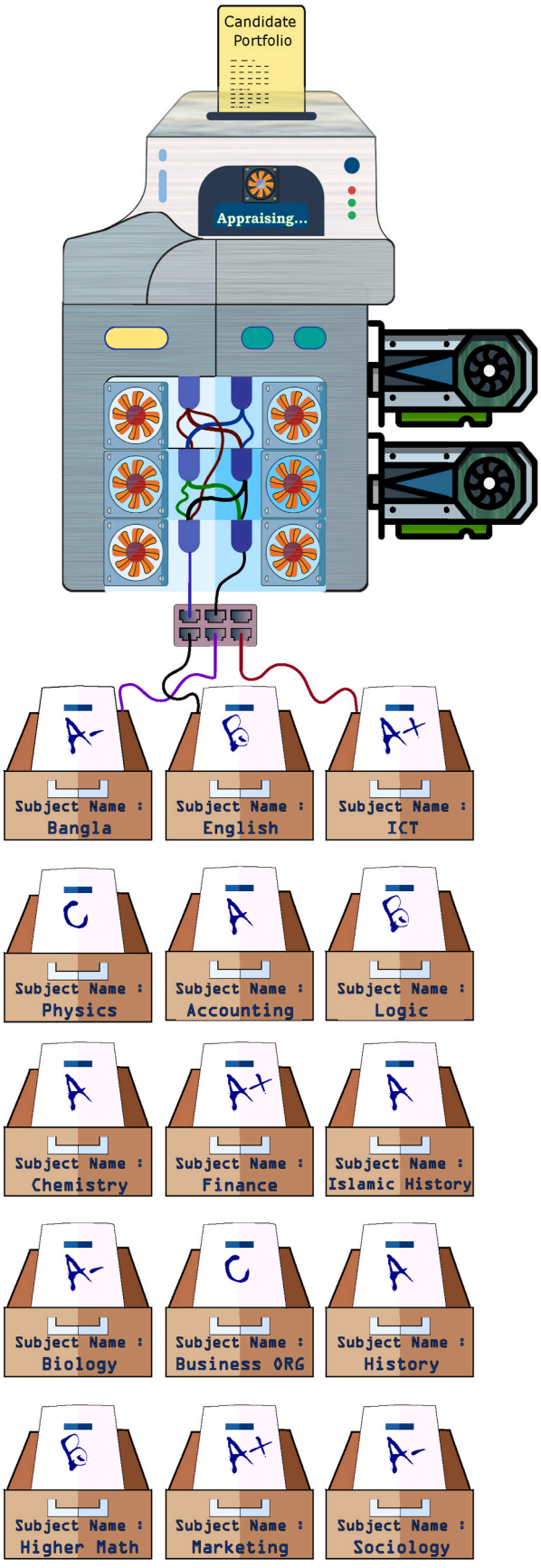


Fig. 1. Proposed hypothetical AI appraisal machine.

model are discussed in 2.6. The research problem, the model's architecture and feature engineering based on students' academic data are analyzed in 3.1, 3.2 and 3.3 respectively. Data are subsequently analyzed in Section 4.1, the model's performance is reported in Section 4.2, and interpretation of predicted scores is provided in Section 4.3. A discussion of the impact of the findings and the study's limitations are given in Section 5. lastly, 6 concludes the paper. The main contributions of the study are summarized below:

- Proposes an alternative inclusive solution to evaluate transcripts of high-stake examinations;
- Harnesses the students' numerical and categorical data to learn the underlying distribution;
- Establishes a generic framework that produces fair and trustworthy evaluations of each candidate of the central education board;
- Develops an autonomous automated evaluation system for Higher Secondary Certificate, which had a detrimental effect on 3.12% of the students only;
- Uses SHAP to minimize the negative impact of including irrelevant features both in the predictive and explanation models.

2. Background & theory

2.1. Researches on academic performance prediction model

In a framework for building an effective student assessment system-the World Bank has reported that the latent assessment strategy can be applied to countries where there is no formal assessment activity or where there is no formal assessment activity or when the education system has been interrupted due to war or other conflict (Clarke, 2011). The world has experienced an educational crisis several times in history. The educational cost due to World War II led to a considerable decline in educational attainment in higher education. The educational disruption due to suspension of enrollment and assessment process becomes apparent if one observes the timeseries data (during, before, after) of the student performance (Ichino & Winter-Ebmer, 2004). As a crisis lasts over time, the odds of educational stress increase correspondingly. The economic loss due to World War II was observed even 40 years after the war (Ichino & Winter-Ebmer, 2004). A trustworthy data-centric method for student assessment without relying on traditional assessment could minimize the economic and emotional impacts by minimizing the dropout rate. In today's world, remote learning tools look practical, but the education sector hesitates to implement an intelligent assessment system for emergencies, despite the fact that sufficient computational resources are available to train predictive models of student academic performance. The purpose of assessment, which is to fairly determine student's progress, can be met using a predictive model as many researchers in the field of AI in Education has suggested. A large-scale project in a primary school in Vietnam has implemented an artificial neural network to predict a student's probability of succeeding in math and Vietnamese (Musso et al., 2020). The model reached very high accuracies (95–100%). Alongside the ability to predict student performance, the model had important implications for policy-makers by highlighting the reasons (the features) for the prediction. A recent study (Rodríguez-Hernández et al., 2021a) has illustrated the most important features contributing to prediction of academic performance. However, the study did not evaluate whether the proposed solution was fair to all learners. In essence, no research suggested an inclusive high-stakes assessment model that is fairer than human decisions and can be considered as an alternative during the absence of examination. It is very challenging to generate a highly accurate assessment model that works for all students and that minimize its detrimental effects. Therefore, more studies should be carried out to develop large-scale assessment models for statewide high-stakes exams.

2.2. Global education crisis

The SDG-Education 2030 Steering Committee provides strategic guidance to the global education community and ensures follow-up and review for the education in the 2030 Sustainable Development Agenda.[5] As soon as Covid-19 spread quickly around the world and was declared a pandemic by the WHO, the Steering Committee underscored that the Covid-19 pandemic was not only a global health crisis but also an educational crisis. The SDG-Education 2030 Steering Committee called on its member states' governments to respect strategic policy recommendations in response to the pandemic (SDG-Education 2030 Steering Committee, 2020). The committee has drawn attention to teachers' and education personnel's safety, health, and well-being. Besides, the Steering Committee urged governments to maintain strong political commitment and investment in education throughout and after the crisis (Guterres, 2020). The UN warned that the pandemic was creating severe disruption in the world's education systems and was threatening a loss of learning, whose impact may stretch beyond one generation of students. The report empirically anticipated the economic impact on households that is likely to widen pre-existing inequities in education. Nearly 23.8 million children and youth (from pre-primary to tertiary) may drop out or not have access to school next year (2021) due to the pandemic's economic impact alone, pushing thereby millions into into severe poverty. Other research by the world bank (Pedro et al., 2021) suggested 25 percent more students may fall below a baseline level of proficiency needed to participate effectively and productively in society. The United Nations encouraged authorities to bring about a set of solutions, previously considered difficult or impossible, to implement and ensure that education systems are more flexible, equitable, and inclusive. UN Secretary-General called for national authorities and the international community to come together to place education at the forefront of recovery agendas and protect investment in education. During this state of confusion and chaos, it's not only pedagogy that will be affected but numerous factors like organizational routines and placement rates at various educational institutions. Both (SDG-Education 2030 Steering Committee, 2020) (Guterres, 2020), foresee that the pandemic will have specifically impacted the education community of low- and middle-income countries. In the future, this may be necessary to conduct a cohort study between the group who has been kept out of school during the pandemic (whose formal assessment was postponed) and the regular group before the pandemic who went through regular assessment. Such studies will uncover the educational, economic, and emotional cost of the Covid-19 pandemic. During 2020, no policy recommendation suggested re-opening of schools or arranging examinations during the emergency. In fact, the United Nations reported that some countries opened schools and colleges, only to close them again after a resurgence of the virus.

International Association of Universities (IAU) is an official partner of UNESCO which acts as the global voice of higher education institutions and organizations from around the world. IAU had planned to carry out three global surveys on the impact of Covid-19 on universities and other higher education institutions. IAU's first Global Survey Report (Marinoni et al., 2020) was conducted in order to capture a description of the worldwide disruption caused by Covid-19 on higher education where the results are analyzed both at the global level and at the regional level in four regions of the world (Africa, the Americas, Asia & Pacific and Europe). According to the survey, 80% of respondents believe that Covid-19 will have an impact on the enrollment numbers of the new academic year. Respondents from Asia & Pacific are the most negative, 85% of them believe that Covid-19 will have a major negative impact on their enrolment numbers since college leaving examinations are on hold and the dropout rate after college might rise.

The World Bank Group has done a rapid assessment of post-secondary education disruption due to Covid-19 (Bank, 2020). The survey suggests flexible adaptations of admission and examination protocols for the incoming academic year to ensure a healthy higher education community during the crisis. Nonetheless, IAU's report has produced motivation for remodeling public policy during the pandemic. Two-thirds of 424 higher educational institutions across the world reported that their senior management and faculty have been consulted by public or government officials in the context of public policies related to Covid-19 (Marinoni et al., 2020). This indicates most current research at higher educational institutions is focused on public policy regarding the Covid-19 pandemic and these researches are being recognized by their respective governments.

2.3. Secondary education system in Bangladesh

Before the partition of India and Pakistan in 1947, the education system of the Indian sub-continent was governed by the British colonials (Rahman et al., 2010). As Bangladesh became a separate independent country from Pakistan in 1971, its education system was restructured under the direction of Dr. Qudrat-e-Khuda (Rahman et al., 2010). Since then, letter grading has been adopted in the assessment of student performance in all phases of secondary school. The higher education in Bangladesh consists of general, technical, engineering, agriculture, business, and medical courses. The minimal criterion for higher education admission is the Higher Secondary Certificate (HSC) or equivalent. The entire secondary education is a seven-year program divided in three stages: 3 years of junior secondary (grade 6–8), 2 years of senior secondary (grade 9–10), and 2 years of higher secondary (grade 11–12). The completion of the three years of junior secondary stage, the two years of senior secondary stage, and the two years of higher secondary stage are assessed by the junior school certificate exam (JSC), secondary school certificate (SSC) and the higher secondary certificate (HSC) test. Fig. 2 illustrates the grade scale from 0 to 5 in X-axis, while the Y-axis represents the passing years. The three horizontal lines in Fig. 2 reflect the three-phase of the secondary school system. The letter grades at the JSC, SSC and HSC level w.r.t the corresponding scores are illustrated in Table 1.

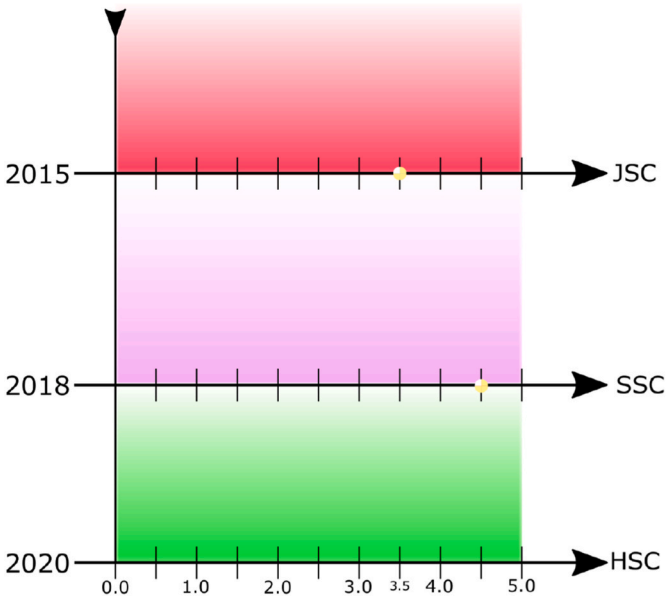


Fig. 2. Grade point range for the Secondary education.

[5] The 2030 Agenda for SDG4 is *"ensure inclusive and equitable quality education and promote lifelong learning opportunities for all."*

Table 1
Mapping of letter grades.

Marks Range	Letter Grade	Grade Point
80–100	A+	5
70–79	A	4
60–69	A-	3.5
50–59	B	3
40–49	C	2
33–39	D	1
0–32	F	0

2.4. Emerging evaluation strategies during emergency

HSC candidates come from all corners of the county, including towns and rural areas. Hence, the college graduation exam is a high-stakes exam as the exam is organized country-wide by the education board. The universal exam-centric assessment method is indispensable for the E−9 countries with large education bodies. During the crisis of the coronavirus pandemic, classroom teaching was hosted online. However, more than one million candidates were waiting to take the HSC exam and proceed to the next stage of life whilst another million from the previous level were added to the queue. As an action to mitigate the educational burden, the Board of Intermediate and Secondary Education of Bangladesh published the HSC result of the year 2020 by simply averaging JSC and SSC grades from the previous academic record (Xinhua, 2020). The model can be understood from the diagram in Fig. 3. Unlike most mathematical models, the structure has parameters $\{w_0, w_1\}$, a functional form $\{average\}$ and variables $\{JSC, SSC\}$ and HSC result as output. In this article, the weighted mean model by the E−9 member country is referred to as the baseline model. According to the baseline model in Fig. 3, the HSC result is an output of the Average function as Equation (1) which takes JSC and SSC result as input. Since every candidate taking the HSC exam has successfully completed JSC and SSC (Educationboard BD, 2020) and the average of two passing grades will also become a passing grade, the model by the baseline model can be interpreted as an auto pass model.

$$HSC_{subject} = avg(w_0 \times JSC_{subject}, w_1 \times SSC_{subject}) \tag{1}$$

where.

$w_0 = 0.25$ and
$w_1 = 0.75$

The baseline model loses the non-linear association of HSC results with other discriminatory variables thereby likely to under-fit the true distribution of real-world observation (Aho et al., 2014). The baseline model, illustrated in Fig. 3, assumes that the real world operates deterministically. This means that the HSC candidate's subject grades tend to occur somewhere between the corresponding JSC and SSC grades. In contrast, this research assumes that the real world is stochastic and that two or more candidates with the same JSC and SSC grades can obtain significantly different HSC grades. The subject-wise formula for All India Senior School Certificate Examination X_f which is the top best performance grade in class X similar to an intercept in a linear equation. Independent of group transition or any irregular event, each subject grade is awarded as a function of X_f which may benefit a large number of candidates while being detrimental for those who could not perform well in class X but deserve better in AISSCE. The $AISSCE_{subject}$ scale is between [0,100].

$$AISSCE_{subject} = \begin{matrix} \sum(w_0 \times X_f, \\ w_1 \times XI_{subject}, \\ w_2 \times XII_{subject}, \\ XII_{practical}) \end{matrix} \tag{2}$$

where.

$f = \sum$(top 3 subject grade),
$w_0 = 0.30$,
$w_1 = 0.30$ and
$w_2 = 0.40$

The AISSCE and HSC assessments were designed considering only regular type of students who have similar performance in both the senior and higher secondary schools. Hence, the baseline model is inadequate to compute the result of all candidates. To address the concerns of the baseline model, human committee was formed to provide recommendations about irregular candidates and more optimal parameter values $\{w_0, w_1\}$ for the baseline model (Hossain, 2020). However, the committee was not allowed to reject the baseline model (Mamun, 2020), that is the functional form must remain as Equation (1).

2.5. Artificial intelligence in academic assessment systems

AI-driven apps are rapidly being employed in different fields of higher education where computer systems automatically analyze digital data and provide recommended actions. A lot of higher educational institutions already merged learning analytics with their existing system. Learning analytics explore the history of students' administrative data, learning activities, etc. to proprovide insights about the past and make predictions about the future. Generally, recommendation systems provide advising services for students as a sort of chatbot (Wang et al., 2021). Chatbots are built using speech recognition and natural language processing. For example, in Germany chatbots assist undergraduate students on their choice of subjects and answer queries regarding courses and course units. It has also become possible to automate admission procedures where an AI model assesses a candidate's portfolio and forecasts if a candidate meets the prerequisites to enroll in a university course. Such a mdoel is trained to understand the pattern between portfolio, the admission decision, and degree completion from historical data, based on past admission decisions made by the human committee. An analogous application is early-warning systems that predict the dropout risk of a student and provide opportunity for educators to help individuals overcome their deficiencies.

As numerous fields of secondary and post-secondary education are adopting autonomous decision-making methods, the reputation of institutions that might use such systems primarily relies on the fairness of the process. Even if AI systems are accurate, they should never violate fair decision-making, legal norms, or ethical principles.

Some input data may unwillingly introduce a prejudice against individuals. For example, an automated classification system in higher education will breach equality if it discriminates against individuals based on their family and religion. In essence, the AI model has to be developed so that it can avoid any infringement of the constitutional norms of the given region. Policy-makers must employ an AI system in higher education only after the risk of losing people's trust and loyalty is prevented. A recent survey among the Dutch adult population (Helberger et al., 2020) found that a greater number of respondents were optimistic regarding the fairness of automated decision making with AI; 54% considered AI a fairer decision-maker than human decision-makers, whereas only 33% believed human-only decisions were fairer than AI. About 9% believed the fairness depends on the context, 3% didn't consider any of the decision-makers fairer than the other, and 1%

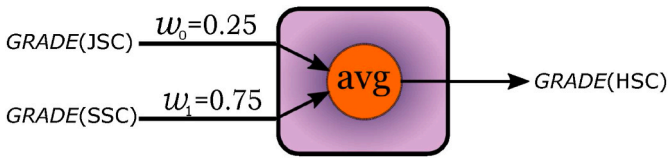


Fig. 3. Conceptual baseline model diagram.

thought both AI and humans should work together.

2.6. Stand-alone components

The first building block of the appraisal model is an artificial neural network. Neural network classifiers with cross-entropy cost function and sufficient sample data can produce outputs as good estimates of Bayesian probabilities (Richard & Lippmann, 1991). Within a national education system, there exists a global underlying distribution of a high-stakes exam grades. Similarly, there exists a grade distribution per district, village, institution, or even course that should also be modeled. For humans, it would take years to understand the pattern of college leaving exam performance with complex historic information stored in millions of text documents. Moreover, it would require a superhuman ability to appraise every student's transcript such that the predicted grades will be matched reasonably with the obtained grades. A universal appraisal framework has to be capable of generating grades that produce a closely similar distribution to the existing one. In the next section, this research investigates the rationale for the appraisal model to critically evaluate the validity of the proposed model. Subsequently, in the following section, the model will be trained with a randomly produced data-set to compare the performance with the implemented model.

Data science and AI systems widely depend on approximation methods. Throughout the years of research, scientists now have access to an Artificial Neural Network which learns any unknown function from a given input space. Not only this powerful tool learns the relationship between input space and outputs, but it also captures the abstraction and distribution of features in the parameter space. Before a machine-learning model was developed in 2016, which consisted of millions of parameters, that finally learned the winning strategy from 150,000 recorded games, it was impossible to find the winning strategy by hard-coding a big set of if-then statements. (Silver et al., 2016). Multiple fully-connected layers of neurons stacked on each other are commonly referred to as Deep Learning architecture in the literature. Topology modifications e.g., residual connections, dropout layers, etc. for some cases, may help the network achieve its objective. Standard deep neural networks are generally effective to approximate any measurable function from one finite-dimensional space to another to any desired degree of accuracy regardless of the squashing function (Hornik, Stinchcombe, & White, 1989); i.e., there exists a single neural network that best approximates the desired function. The universality theorem shows promising results in practice which is possible to implement due to recent advances in hardware designed for deep learning. To mention a few real-world use cases, Frontier Development Lab (FDL) has developed a method, based on deep learning, that allows researchers to more accurately capture the Sun's extreme ultraviolet irradiance (Szenicer et al., 2019). Extreme ultraviolet irradiance is responsible for solar flares — has, in the past, typically been monitored by recording the distribution of magnetic fields or plasma on the sun and feeding that data into a physics-based model to predict EUV emission. The results of the deep learning approach surpass current physics-based models – and, as part of the process of evaluating their results, the team developed new benchmarks for comparing predictions. Another deep learning-based research breakthrough of Frontier Development Lab in accomplishing the first large-scale simulated spectral data set and first atmospheric retrieval ML model for rocky, terrestrial exoplanets by introducing INARA (Intelligent exoplaNet Atmospheric RetrievAl) (OBeirne et al., 2019). Moreover, at present and in the future, deep learning an AI applications are more likely to be used in educational sectors (Zhang & Aslan, 2021). Recently, deep learning models have been demonstrated to be a valuable tool for detecting students' engagement (Dewan et al., 2019), recommending next learning activity (Dewan et al., 2016), automating essay scoring (Kumar & Boulanger, 2020) and many pedagogical applications. Deep learning is a powerful tool in terms of extracting pattern from transformed input features through which it learns complex relations and makes predictions with high precision. Each layer makes abstraction of given representation from high level to low level proportionally with the depth of the model. In essence, the promising performance of neural networks is achievable because of the ability to transform input space (Bengio et al., 2013) into the latent space. Whether the task is to determine output(s) from input space (deterministic model) or generate entire input space from partial inputs (generative model), neural architecture proves robust in all cases with the power of representational learning. Deep learning offers great flexibility in the way to design and adapt a neural network to a domain-specific problem. For example, a generative model which requires reconstruction of the input(s) can have bi-directional connection edges. Such a model having bi-directional edges with a bipartite architecture (one input layer & one hidden layer) was proposed in (Salakhutdinov et al., 2007) as Restricted Boltzmann Machine (RBM). An RBM with an arbitrarily large number of neurons has been demonstrated as a universal model for complex data distribution (Larochelle & Bengio, 2008). Stacking several RBM, a deep learning model can be created and in many cases, the topology outperformed the typical feed-forward network. For instance (Hassan et al., 2019), extracts abstractions of physiological signals through three stacked RBM to find the complex relationship of human emotion with the input signals. Conditioned on independent variables, the binary outcomes are produced using a sigmoid function in logistic regression (Bonney, 1987, pp. 951–973). A logistic regression model projects the independent variables into a one-dimensional space which goes into the squashing function to produce binary outcomes. Multinomial logistic regression modeling is suggested over statistical modeling to identify anomaly intrusion (Wang, 2005).

AI model interpretability is indispensable to understand how each feature contributes to the final outcome. However, machine learning models with a complex architecture cannot explain its predictions. Therefore, the SHAP (SHapley Additive exPlainer) framework was used to address this tension between accuracy and interpretability (Lundberg & Lee, 2017). SHAP measures feature importance using the same scale as the one for predictions, which makes it easier for humans to interpret. A perceptive (Tjoa & Guan, 2020) interpretability framework has more usage in the field of medical AI or high-stakes decision-making AI to achieve responsible AI. The mean Shapley values or base values of features can be useful to interpret variable importance for all data points (Bosch, 2021).

3. Methodology

3.1. Problem setting

The Turing test is a popular method to determine a machine's ability to exhibit intelligent behavior. Alan Turing in his evolutionary paper (Turing, 1950) first coined the concept of the Turing test in terms of an imitation game which is a quantified approach in the quest to determine whether a machine can think. There are three entities in the imitation game: an interrogator and two participants X and Y, where one is a human participant and the other is a machine participant. X and Y both perform an activity as told by the interrogator, such as attempting a math problem, cooking and serving a dish, performing a medical diagnosis, writing computer programs, summarizing articles, driving a vehicle, etc. The objective of the machine is to perform the assigned task with adequate intelligence such that its outcome is barely distinguishable from the human one. In the case of an intelligent machine, an interrogator will not have more than a 70% chance of making the right identification. The exam grade appraiser model in this research is considered as the machine in the Turing test using the same game setting described by (Turing, 1950). In essence, upon inserting the student identifier, the associated transcript of the high-stakes examination will be shown to the interrogator's output device. The interrogator will receive two transcripts, one produced by a machine another achieved from the human evaluation process. The interrogator will see the retrieved grade sheet for the first time and will have zero knowledge on

how the student performed in the submitted exam. However, the interrogator can access the previous portfolios of students and the distribution of student features. The proposed model is designed with the hypothesis that it can produce outcomes similar to what is exhibited in the real world. Therefore, the proposed model is trained to imitate the transcript generated from the real world where students take an exam. To persuade the interrogator, transcripts from the sources X and Y have to be very similar so that they look as coming from the same universe. In contrast, a large dissimilarity will make the interrogator believe that one of the transcripts is generated by a rule-based system.

The appraisal machine's endeavor is mathematically described in the objective function in Equation (3).

$$L(\theta) = \frac{1}{m}\left\{\sum_{i=1}^{m}\sum_{c=1}^{k} -y_{c,i}\cdot\log\,\widehat{y_{c,i}}\right\} \tag{3}$$

where the output for each course is a one-hot encoded vector. The true labels and the predictions $y_{c,i}$, $\widehat{y_{c,i}}$.

$$y_{c,i} = \begin{bmatrix} 0.0 \\ 1.0 \\ \vdots \\ 0.0 \end{bmatrix} and, \widehat{y_{c,i}} = \begin{bmatrix} 0.0 \\ 1.0 \\ \vdots \\ 0.0 \end{bmatrix}$$

m denotes the number of candidates and k is the number of courses taken by each candidate. A machine that outputs a constant grade for all students will be easily distinguishable as a machine-generated transcript. Similarly, simple arithmetic operations such as averaging previous grades, maximum/median/mode of institutional performance, or as a function of junior and secondary school grades will first and foremost make the two transcripts in front of the interrogator highly dissimilar and secondly, due to the fact that interrogator has access to the information of portfolios and distributions, therefore, such assessment formula will fail the Turing test.

More than one million (13,65,789) candidates for the HSC exam for the 2020 session have been assessed by the baseline model. As opposed to the baseline model, the proposed robust model learns underlying distribution on its own from a complex representation of quantitative and qualitative features, therefore, guaranteeing maximum fairness in evaluation. The quantitative features include each candidate's obtained grades in junior secondary and higher secondary level; apparently, every candidate appearing higher secondary exam previously ensures[6] marks in all subjects of JSC and SSC within the range of [33,100]. Therefore, presumably, no portfolio should contain fail grades in mandatory subjects of the JSC and SSC examination given that the candidate is an HSC examinee. The Indian government's AISSCE assessment formula as well as the HSC exam assessment formula both exhibit pass grades as an output of the appraisal process for being consonance with past performance at the school level. The proposed model learns higher-level distribution of the achieved grades as well as low-level distribution e.g., specific education group, board, institution, and so on. It is not required to provide external bias or any heuristic about underlying distribution for real-world to the model, it is the intelligence of the robust architecture to make a trade-off in generalization and precision in order to capture the distribution only from the real-world data. The human task is to investigate how well the machine learning model fits the underlying distribution and whether it learns the complex relationship of HSC grades with given features with a certain degree of complexity. In order to train and probe the machine learning model, the empirical data is obtained from the previous three sessions of the 2020 session that is denoted as, $m \in \{(x^1, y^1), (x^2, y^2), \ldots, (x^m, y^m)\}$. Before delving into the architecture, notations used in the framework are arranged in Table 2.

The notation x^i is the portfolio feature representation of the HSC candidate as a generic symbol to represent all quantitative and qualitative information to feed into the model. Along with various significant qualitative information, geographical location is considered in the feature space by incorporating associate institute's EIIN number[7] as a qualitative categorical variable. A 6 digit EIIN is a unique ID of institutions and therefore, used as an encoded information about an institution's location. A more detailed overview of feature construction is mentioned in Section 3.3.

Table 2
Symbols used in the proposed algorithm.

$id \subseteq \mathbb{R}^{D_x}$	unique identifier of each HSC candidate
$x^i \subseteq \mathbb{R}^n$	representation of an HSC candidate associated with ith id
$\theta \subset \mathbb{R}^d$	parameter space
$L(\theta)$	objective function to minimize by the model
T: $2 \times 7 \rightarrow [0-1]$	educational group transition vectors
$e_i \subseteq \mathbb{R}^{10}$	indicator vector to represent single bit of EIIN
$E^i = (e_{i_{MSB}}, \ldots, e_{i_{LSB}})$	EIIN encoding vector by stacking indicator vectors
$B^i \subseteq \mathbb{R}^{11}$	indicator vector of position 0-9
$e^{i_{nr}} \subseteq \mathbb{R}^3$	indicator vector that encodes the category- regular, improvement or retake for $x^{(i)}$
$v^i \in [0-1]$	binary bit represents either Bangla or English version in HSC level
$\varepsilon \in \mathbb{R}$	step size in gradient descent
$y^i \subseteq \mathbb{R}^7$	HSC marks of 7 subjects associated with x^i
$m \in (\mathbf{X}, \mathbf{Y})$	size of data sample or number of candidates

3.2. Overview of appraisal Model's architecture

A complex model with many parameters is proven to fit data better than a simple model with few parameters (Myung, 2003). Therefore, unlike the baseline model which contains a small number of parameters, the proposed machine learning pipeline contains nearly a million model weights denoted as θ and initially, the model parameters do not obtain any prior distributional assumption about the true distribution (see Fig. 4). Model outputs are referred to using the y-hat notation $\hat{y}$ generated upon providing input arguments (θ, x) to the appraisal model. Following a tweak or dynamic modification of the model parameters, a different probability distribution is generated each time (Myung, 2003, eq. (1)). The appraisal model is allowed to tweak its θ as long as the model moves closer to zero error. The appraisal model finds the adjustment of the θ aiming to reduce cost function while fitting the underlying distribution of the provided high-stakes exam performance. Given m number of candidates of previous sessions, there are feature vector $\mathbf{X} = (x^1, x^2, \ldots, x^m)$ and observed output vector $\mathbf{Y} = (y^1, y^2, \ldots, y^m)$. A random sample from the population of past candidates is given to the model during the training or, tweaking phase to adjust its θ as such the model produces a probability distribution $f_p(y|x, \theta)$ which is most likely the underlying true distribution. The optimal parameter searching during the tweaking phase is formulated as Maximum Likelihood Estimation (Aldrich, 1997). Appraisal model for the nation-wide exam candidates learns to appraise with only one objective function that expresses the deviation in appraisals from the actual outcome. No prior bias for any particular group of candidates is given as such the appraisal model loses its validity.

For categorical outcomes, the objective function for the appraisal model is the categorical cross-entropy function provided in Equation (3). Several choices of the objective function $L(\theta)$ are provided in Equations (4) and (5) specifically for numeric outcomes. Equation (5) allows the policy-makers to set a penalty term as hyper-parameter. Policy-makers can decide to add on a non-detrimental guarantee using a penalized objective function as Equation (5) where the penalty factor $\gamma \in (0.5, 1.0]$

[6] Rules for Secondary Education in Bangladesh.

[7] Unique identifier in a government database as "Education Institution Identification Number"EIIN.

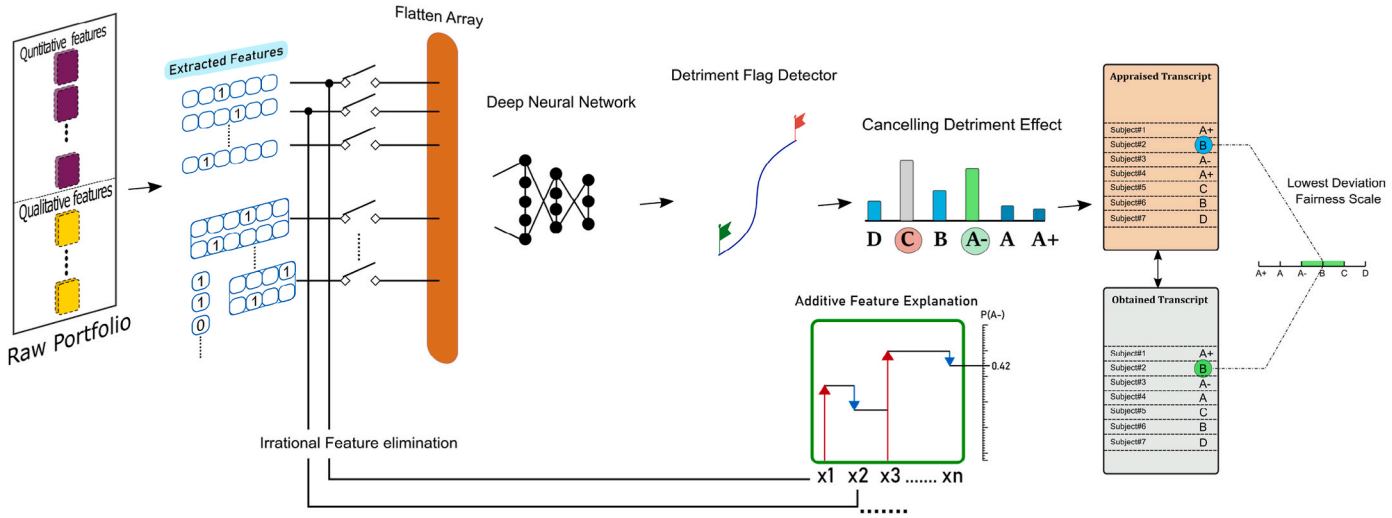


Fig. 4. Schematic diagram of the proposed method. Raw portfolio features are extracted as one hot encoding vectors (for categorical features), then a deep neural network receives the flatten array as the extracted features. By backpropagating log loss function, the required change of θ is accomplished using gradient descent technique until global minima of the log loss is achieved. Events of detriment appraisal is stored as training dataset for logistic detriment classifier. Only a detection of detriment effect is required for the modification of the output distribution. Once θ is converged, manually eliminating a feature from flatten array is allowed before publishing the grade on transcript. Feature contributions are revealed with Additive feature Explainer.

protects the model to output distributions that are below the true distribution. For equiprobable outcomes, the γ will motivate the appraisal model to assign the higher marks which might eliminate the risk of detrimental effect.

$$L(\theta) = \sum_{i=1}^{m} (y^i - \hat{y}^i)^2 \tag{4}$$

$$L(\theta) = \sum_{i=1}^{m} \left\{ \sum_{i=y^i<\hat{y^i}} (1-\gamma) \times |y^i - \hat{y}^i| + \sum_{i=y^i>\hat{y^i}} (\gamma) \times |y^i - \hat{y}^i| \right\} \tag{5}$$

On the other hand, while appraising irregular candidates, their portfolio might contain several subject grades of the HSC examination. Hence, the past history of HSC is provided along with the portfolio for appraisal of the remaining HSC grades. In terms of appraising irregular candidates, the aim is to appraise some or all HSC subject grades such that the fairness of evaluation is maximized by appraising grades which are more likely the observed real results. Now, to determine whether a generated output by the model is close to the true distribution of previous graduates', an Energy state is introduced (in Equation (6), a joint configuration of the portfolio, HSC grades, and a layer of neurons) between the irregular candidates' portfolio and output labels. A low energy state close to zero is expected to achieve similar outcome as the real-world data.

$$E(y,h,x) = -y^T W^{yh} h - y^T b^y - x^T W^x hh - h^T b^h \tag{6}$$

The model can approximate any true distribution p by approximating q and the precision of approximation w.r.t true distribution is their difference measured as relative entropy in Equation (7).

$$KL(p\|q) = \sum_{x\in X, y\in Y} p(y|x) \ln p(y|x) - \sum_{x\in X, y\in Y} p(y|x) \ln q(y|x) \tag{7}$$

where $KL(p\|q)$ is zero when $q = p$, otherwise it is non-negative. By going downhill of the engergy function, the model seeks to yield $KL(p\|q)$ to reach to zero. The aim is therefore, to find the optimum configuration of θ in Equation (8) that reduces the gap between the energy of the observed data and the produced data.

$$\theta^* = \arg \min_{\theta} \sum_{y\in Y_p} E(y,h,x) - \sum_{y\in Y_q} E(y,h,x) \tag{8}$$

The objective function $L(\theta)$ for irregular candidates deals with a free energy function associated with the Equations (6) and (9).

$$F(y,x) = -\log \sum_{h} \exp(-E(y,h,x)) \tag{9}$$

Finally, the objective function when the partial outcomes are known, is defined as in Equation (10) which was introduced in Conditional Restricted Boltzman Machine (Mnih et al., 2011):

$$L(y,x) = F(y,x) - F(\hat{y},x) \tag{10}$$

The core unit of the proposed architecture is the artificial neuron, which is also the building block for the human brain. Billions of neurons form a human brain that is capable of multiple tasks. In this context, the artificial brain is specifically designed to meet the goal of the identified research objective. The arithmetic block is loosely a resemblance of brain neurons, which flashes activation signals; such type of arithmetic block is largely used in solving reasoning tasks. Here, a more low-level conceptual diagram of arithmetic block is shown at the amplified section in Fig. 5.

A Layer of neurons is stacked together without any inter-connectivity among them. Therefore, artificial neurons are the only processing block as the backbone of the proposed appraisal model. Here, the core framework is composed of two pipeline blocks, Block 1 and Block 2 as shown in Fig. 5. Blocks are enabled based on a candidate's registration type using logical operation in Fig. 6. The difference between Block 1 and Block 2 is that Block 2 processes portfolios where partial information from HSC or AISSCE exams are known. Block 1 processes information of the candidates who are first-timer of the high-stakes exam, where as block 2 processes the irregular candidates who have previously appeared in the exam and retaking a few courses or reappearing. The first pipeline is feed-forward in nature, which means Block 1 does not receive HSC marks as input. In contrast, the second pipeline has fully connected bidirectional links with HSC marks. The Block 1 and Block 2

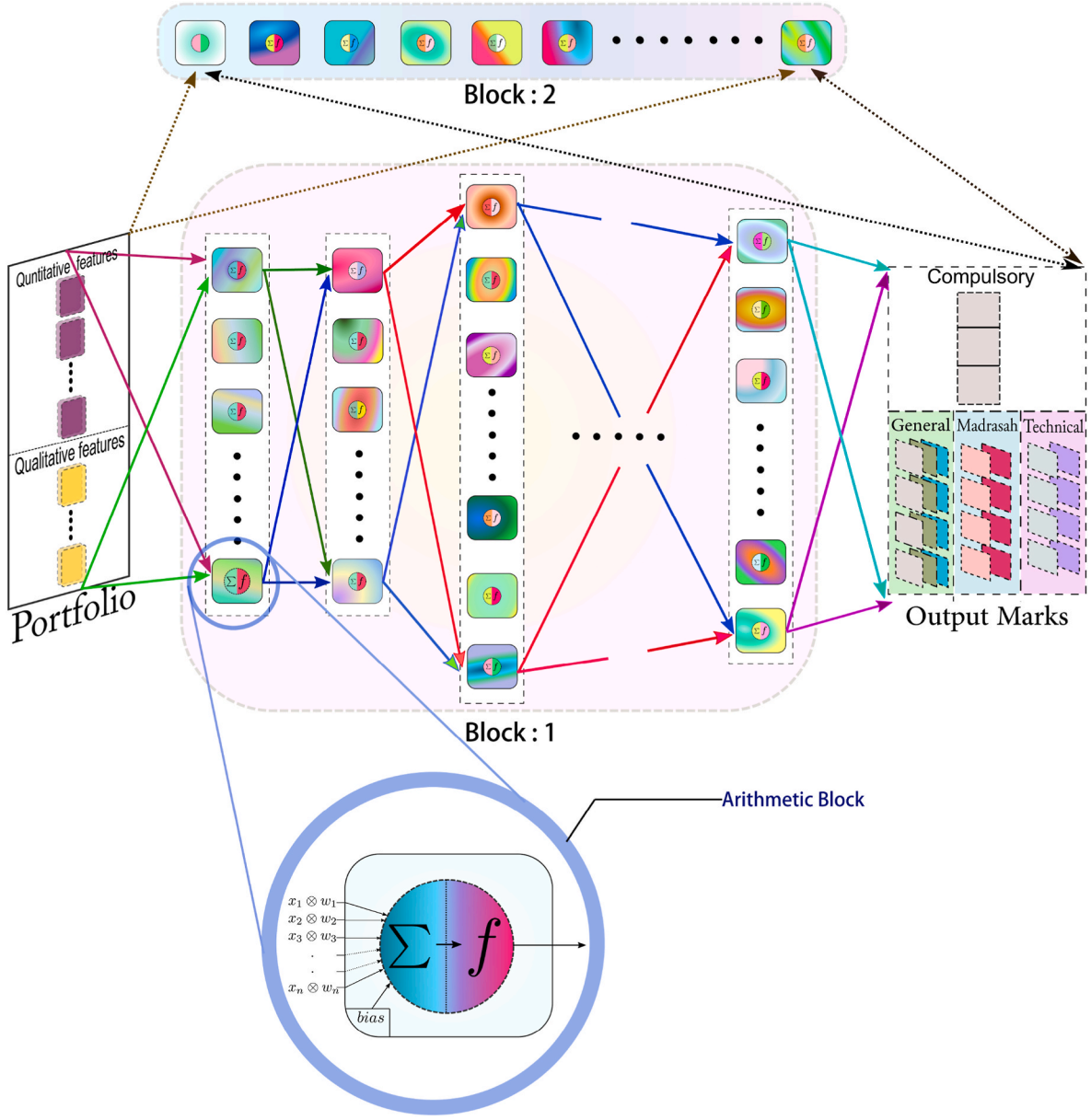


Fig. 5. High level architecture of Deep Learning Model.

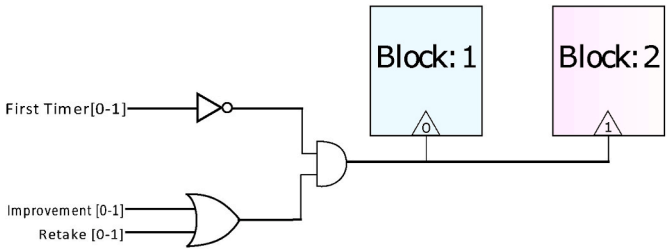


Fig. 6. Model pipeline control logical unit.

processes are feed-forward deep neural networks and restricted Boltzmann machines, respectively.

Each neuron of the backbone model is composed of net input function $\sum(x_i w_i + b)$ and a non-linear activation function $f(\sum(x_i w_i + b))$. Assigning a non-linear f for the appraisal framework substantially depends on empirical analysis from the widely acknowledged set of activation functions. State-of-the-art activation functions in practice is thoroughly reviewed in (Nwankpa et al., 2018, Table 2). In the context of the HSC appraisal model in the empirical analysis, Rectifier Linear Unit produced optimal outcomes. The model will output individual subject marks and the "Marks Output" layer does not have any activation function. As a contextual model, Fig. 5' output layer is illustrated with all the possible subjects at the HSC exam in the output block. The HSC exam transcript generally has seven subjects where three subjects are compulsory and the other 4 subjects are dependent on a candidate's educational group. In the case of categorical letter grades, the deep learning output layer would be referred to as Output Letter Grades instead of Output Marks. To output letter grades, each subject will have as many nodes as the number of letter grades. Thus, the appraisal model's outcome is multiple headed multi labels for categorical grades. Traditional binary output neural network approaches (Rodríguez-Hernández et al., 2021b) in small-scale university level performance predictors are not intended to produce subject-wise non-detriment appraisal rather can be used to weakly estimate a positive or negative change in future performance. The process allows the deep learning model to make the trade-off with detriment effect to achieve

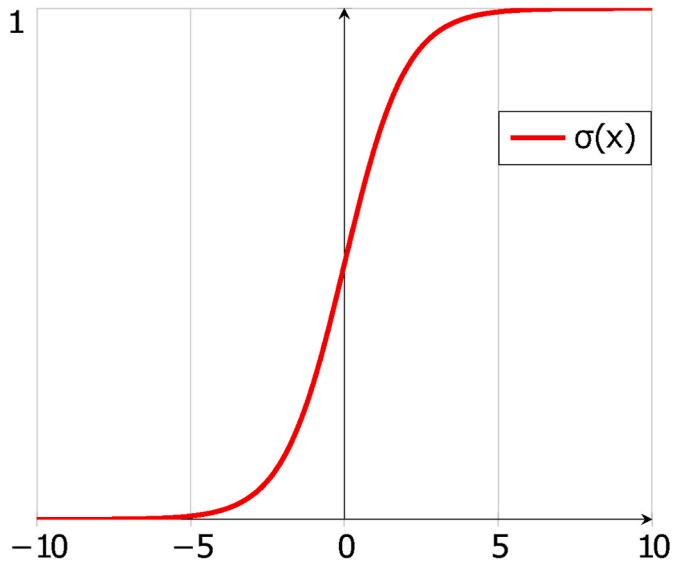


Fig. 7. Sigmoid function for logistic regression.

maximum fairness level. The detrimental flags are labeled and passed into logistic regression models to classify detriment and non-detriment appraisal. The sigmoid function in Equation (11) takes the portfolio information, Output layer of backbone model, and probability distribution's entropy (for multiple categorical outputs)- then squash the features into the X-dimension shown in Fig. 7.

$$\sigma(x_i w_i + b) = \frac{1}{1 + e^{-(x_i w_i + b)}} \tag{11}$$

Multiple logistic regression models were stacked for as many multi head output. As the probability distribution of the output layers is modified once a logistic regression classifier detects a detriment flag, the assigned grade by the backbone deep learning model is discarded by considering that the predicted grade has higher chance of being a detrimental assessment. Then the autonomous evaluation process will again predict a higher grade than the previously appraised grade. In the later process, some of the input factors may be required to multiply with zero in order to disable the effect of those particular features. An example of the outcome probability correction after detriment flag detection is shown in Fig. 8 where the subject grade was awarded a letter grade 'C' by the backbone model. Following the detriment flag detection, the 2nd highest higher grade 'A-' is chosen as the output grade. Once the iterative correction process is done, the outcomes can then be generated applying the soft-max function on the distinct probabilities.

The appraisal interpretability by SHapley Additive feature value is obtained by the kernel SHAP algorithm that observes the contribution of a single feature assuming all features are independent of each other. The explainer is a vital element of the process to achieve zero detriment policy.

3.3. Feature construction

From the facts and the data exploration, it can be observed that a candidate's portfolio of higher secondary examination in Bangladesh must inherit one category from each branch of the qualitative feature space as shown in Fig. 9. The members of a branch are mutually exclusive.

Due to nominal categorical property and mutual exclusivity, the qualitative features are quantified using an indicator vector. At a high level in the E−9 member country's higher secondary education system, three education boards are - General, Madrasah, and Vocational or Technical board. Based on geographical location there are 11 educational boards in the country. Except for the language courses, the rest of the syllabuses are either in Bangla or English version which is also considered as a distinguishable feature, because the colleges generally have two separate wings based on the educational versions with independent management and infrastructures. Educational group is another significant criterion based on educational background. The education group determines a set of subjects that a student must pass to earn the certificate. Moreover, the Higher secondary level education group is coupled with information about group change after school by

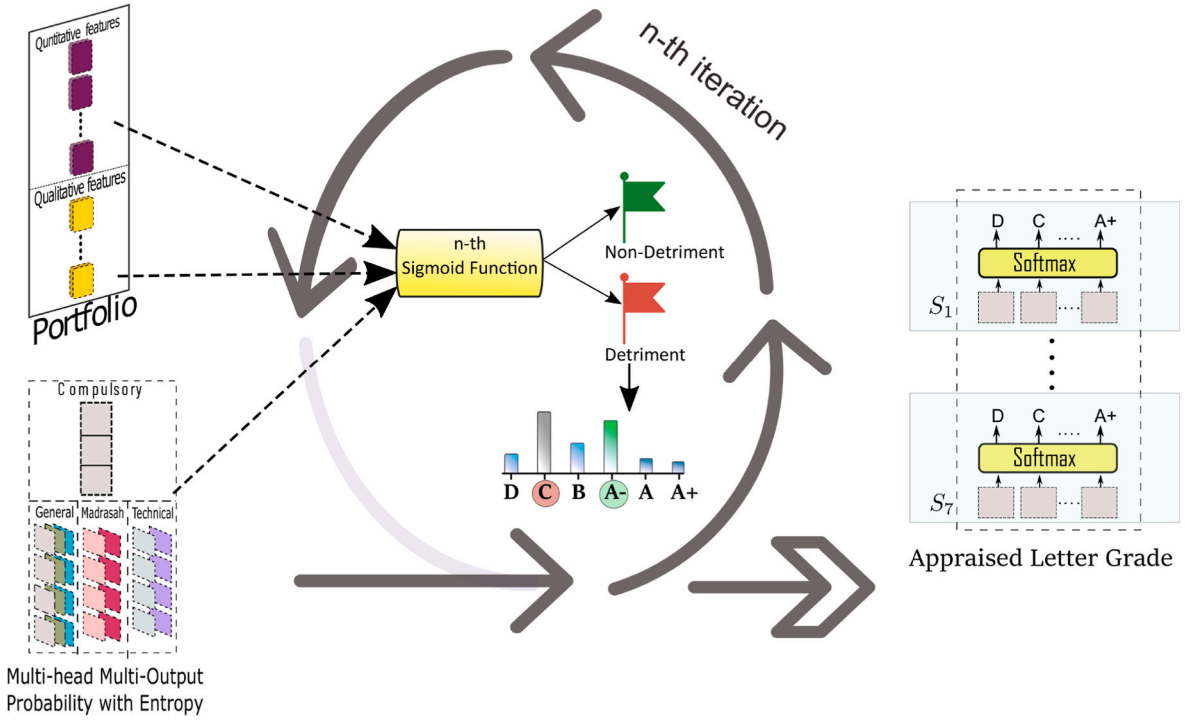


Fig. 8. Iterative Process of Detriment Flag Detection before appraisal of transcript grades.

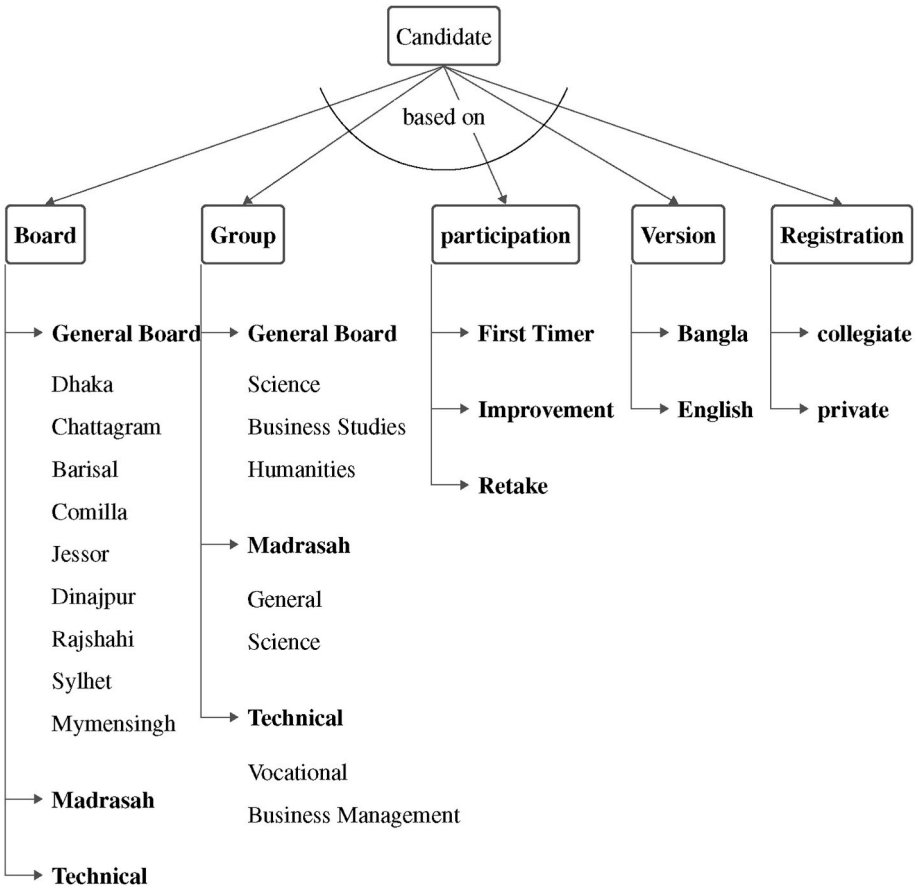


Fig. 9. Categorical attribute tree of senior secondary board candidates in Bangladesh.

considering a two-dimensional matrix. Existing appraisal models are unable to address the education group history whereas this parameter is not only is a significant factor in the higher secondary examination result but also vital to design a rationale appraiser. What was a candidate's group at secondary school level and at the higher secondary level is represented as the following (2×7) matrix.$T = \begin{matrix} G.science & G.commerce & G.humanitics & M.General & M.Science & T.Business & T.Vocational \end{matrix} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{matrix} SEC \\ HSC \end{matrix}$

A candidate can be registered to take the public exam from an institute or privately. Students who could not pass in one or more subjects in the previous session re-register in the following year for improvement or retake the entire test. Categories with binary labels, e.g., educational version, registration type, etc. are encoded into a single binary bit where 0 and 1 represents distinct labels of specific category as shown in Equation (12).

$$X = \begin{cases} 0, & \text{if } a = 1 \\ 1, & \text{otherwise} \end{cases} \tag{12}$$

Particularly the qualitative feature space is initially selected by conducting a student's t-test when a significant P-value is found (Brevard & Ricketts, 1996). Recent studies have shown that the prior academic achievement, socioeconomic conditions, and high school characteristics are important predictors for academic performance (Rodríguez-Hernández et al., 2021b) (Rizvi et al., 2019). Besides qualitative properties mentioned in Fig. 9, topographical information and particularly the migration event is discovered as an impacting variable due to significant P-value. This feature did not require additional pre-processing to incorporate because a unique educational institution identifier or EIIN is a 6 digit number that exists with JSC, SSC, and HSC transcripts. The postal code, village, road, district, etc. topographical information are encoded within the first 5 digits of the EIIN number. Since the EIIN code does not represent a numeric value hence, the five-digit decimal code is converted into a (10×5) matrix $\boldsymbol{E}$ where columns are indicator vectors in Equation (13).

$$E^i = \begin{bmatrix} e_{0,5} & e_{0,4} & e_{0,3} & e_{0,2} & e_{0,1} \\ e_{1,5} & e_{1,4} & e_{1,3} & e_{1,2} & e_{1,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ e_{9,5} & e_{9,4} & e_{9,3} & e_{9,2} & e_{9,1} \end{bmatrix} \tag{13}$$

However, privately registered candidates in Fig. 9 do not appear in the exam from any institutions. An option to make the model understand a candidate is privately registered is to use a custom predefined constant EIIN e.g. EIIN: 99999, which is non-existing and can be represented with the matrix $\boldsymbol{E}$. Finally, other variables such as gender, age, family members, and place of birth did not show significant impact during the T-test hence, those features are eliminated from the student portfolio.

The mapping of grade points is shown in Table 1. The baseline model discards the "Fail" grade and the proposed model does that as well by considering six passing grades as the output. Quantitative representations of subject-wise marks from JSC and SSC are clipped between the range $[33, 100]$ and the subject-wise output marks of the higher secondary examination are also clipped between the range of pass marks$[33, 100]$. While training the appraisal model with the aim of appraising numeric scores, it is efficient to scale down the range in between $[0, 1]$ using the equation in Equation (14) in order to achieve faster convergence (Wan, 2019).

$$x_{i_{norm}} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{14}$$

The final result sheet of the HSC examination carries letter grades of seven subjects along with a cumulative grade point average based on a

scale that is generated from the range of total obtained marks out of hundred. Since the Higher Secondary Examination transcripts in Bangladesh contain letter grades shown in Table 1 therefore, in this research JSC and SSC subject grades are provided in ordinal scales as well as the labels produced by the appraisal model are also categorical. Thus, each candidate is a data point in high dimensions derived from qualitative and quantitative features from the candidate portfolio.

The training data of this study has zero (1) erroneous attribute values, (2) missing attribute values, (3) incomplete attributes values. Inevitably, the training data got rid of the sources of "attribute noise" (Zhu & Wu, 2004). The two other sources responsible for "class noise" are-wrongly labeled instances and contradictory examples (Zhu & Wu, 2004). Each training instance of the real-world data is validated by many stakeholders before being published as a record. Moreover, In the event of a probable mistake in the published grade, the training data has also gone through the student inquiry procedure in the real world. Thus, the chances of a mislabeled instance are negligible. A high rate of contradictory examples can often cause the appraisal model to have fallacious decision boundaries. One common issue in the classification task is the noise of training samples that may lead to the creation of small clusters of one target class in the region of the domain corresponding to another class. If a lot of such data points are near the decision boundaries then may create a hindrance for the learning algorithm and increase the complexity due to over-fitting behavior (Libralon, de Leon Ferreira, & Lorena, 2009). Fortunately, neural networks try to fit the noisy data after many iterations of the gradient descent. During the initial training process, the model only focuses on the true signal. Applying an early stopping strategy to the appraisal model will provably help the model to learn from true signal (Kingma & Ba, 2015). Because stopping the learning process before it over-fits the noisy samples will prevent the model to memorize all the boundary data points. Early stopping was set to be activated when validation performance started to decrease even though training performance was going up. Furthermore, pruning the noisy data points during the training process will help the appraisal model to learn the task more confidently (Northcutt et al., 2021). Therefore, early stopping (Kingma & Ba, 2015) and confident learning (Northcutt et al., 2021) was applied to suppress the effect of noise during the training process.

4. Performance evaluation

Although the performance is evaluated based on model prediction the accuracy of the appraisal process highly depends on the correctness of training samples the model receives. Therefore, it is indispensable to first ensure (*I*) the typical distribution of class labels across training and testing observations and, (*II*) the true rate of signals in the machine learning data. The first concern about the class distribution is discussed in 4.1. Lastly, this study will answer the following issues in 4.3:

- If the training data had useful signal in it.
- If the number of training sample was useful for training.

4.1. Data Descriptiom

The portfolio data is provided to the appraisal model during learning along with the HSC transcript from the passing years of 2017 and 2018. This research includes over one million (1015617) anonymized observations from three consecutive years—2017, 2018 and 2019—to produce a like-for-like comparison. The data set contains candidates from all three study groups: science (31%), business studies (28%), and humanities (39%). Fig. 10 depicts the year-by-year letter grade distribution of each subject. The distribution curves for grades follow an underlying distribution with no significant disharmony. For example, the English course has been following a long-tail distribution where getting an A+ has the lowest frequency (2.54–2.85%) and each subsequent grade is a more probable outcome. The goal of inspecting the underlying distributions in Fig. 10 is to determine whether the result distributions of subsequent years are related to previous distributions.Relative entropy, i.e., the difference between the distribution of class labels in the train set and the test set, is much lower than suggested, considering the previous 2 years as the learning period for the model and the appraisal model.

4.2. Model performance

Given the portfolios from the training set, the model learns to match each candidate's HSC grade such that the categorical cross entropy loss is reduced. The subject-wise confusion matrices exhibit what portion of each grade point (row) is appraised as the predicted grades (column). The confusion matrices in Fig. 11 are row-wise normalized.

The linear trend of higher values along with the main diagonal show machine learning outcomes almost match the true grades for new candidates of the year 2019. A weak model would not be able to capture the complex pattern in the portfolio and would appraise randomly, or one output would be given for all candidates. Notably, the appraisal model passes the abnormal or irrational prediction test with the confusion matrices. There is no such case of appraising only the D grade (or any arbitrary grade) for all the candidates or randomness in the appraisal. As a responsible AI model, it is indispensable to make sure that, in the worst case, only a portion of candidates are treated inequitably due to erroneous appraisal. In addition to the confusion matrix to scrutinize appraisal model performance, the maximum chance of being a loser due to machine error is sought in each subject using the scale in Fig. 12. The heat map in Fig. 13 shows the portion of losers per hundred is significantly low. The heat map is sort of a litmus test for the responsible appraisal model, which will help researchers benchmark their model in this domain. Any rational alternative appraisal model will have some reasoning errors, as the real world is non-stationary. A model with no errors will achieve the highest value of 1.0 at Fair-Zone in Fig. 13. As the unified appraisal model (one architecture and objective function) introduces some deviation for the test data set in Fig. 11, the litmus test is conducted by comparing each appraisal with truly obtained grades in an individual subject using the deviation scale in Fig. 12. In the real world, two identical deserving candidates will frequently have a one-letter

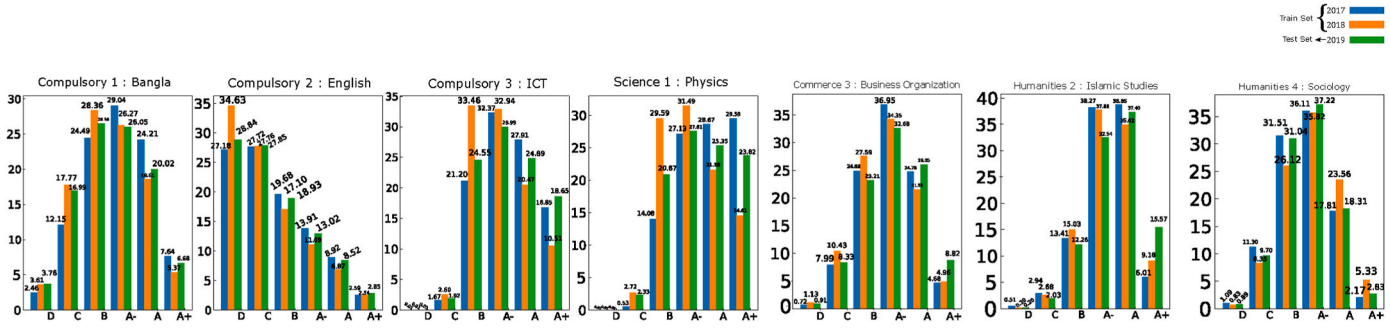
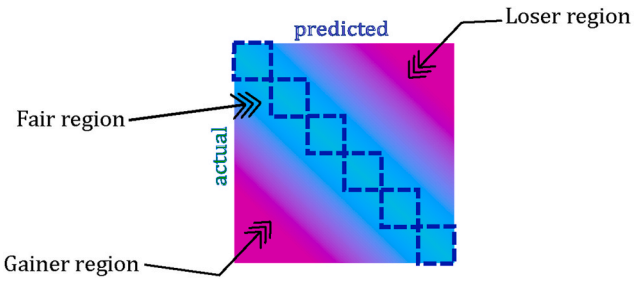


Fig. 10. Year-wise Grade Distributions of each subject.



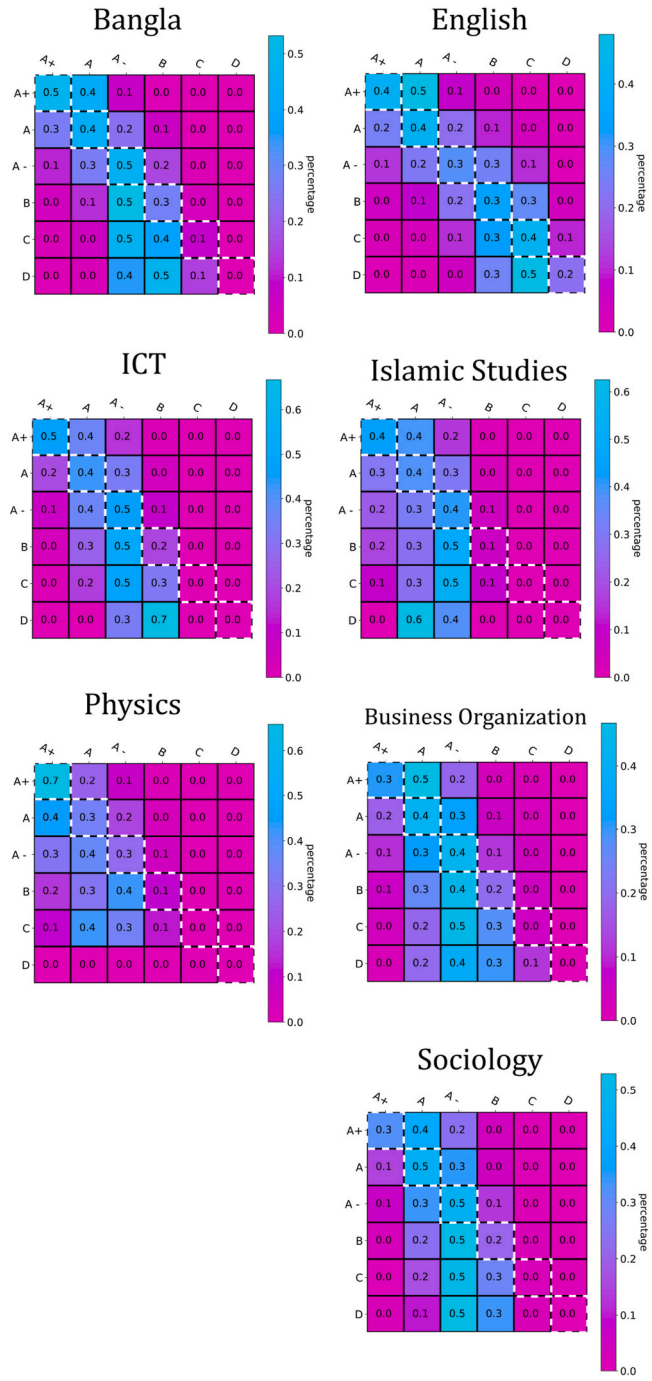


Fig. 11. Confusion Matrix for subject-wise appraisal among test session candidates.

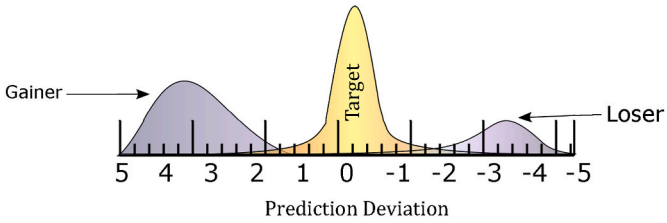


Fig. 12. Fairness, gainer and loser scale for quantitative evaluation.

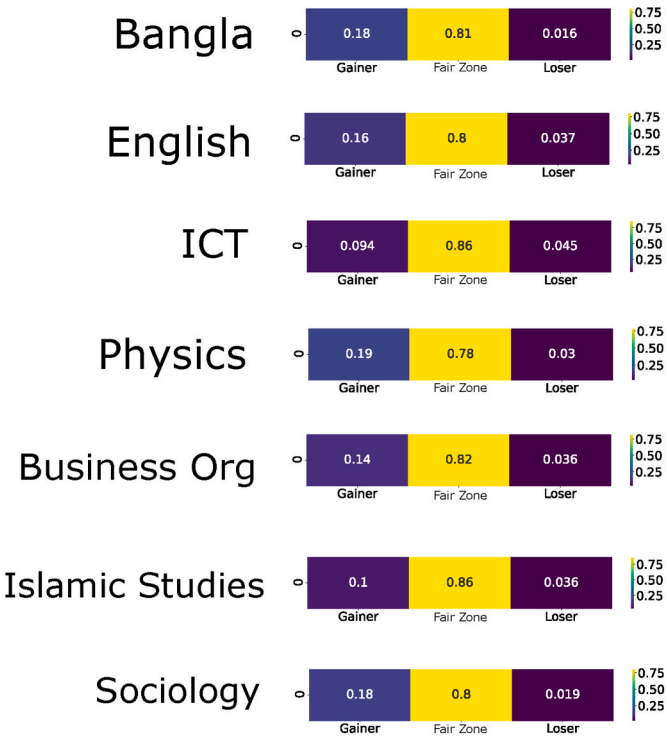


Fig. 13. Gainer, Fair and Loser percentage among test session candidates.

grade difference in their transcript due to a 1 or 2 score difference in the exam script. At most one neighboring grade point is therefore considered as a fairness boundary since the output targets have orders, e.g., A and A- are adjucent, whereas A and D are largely dissimilar. ***If***, ***G*** denotes a set of letter grades then, $\boldsymbol{G} = \backslash\{(A+), (A), (A-), (B), (C), (D)\backslash\}$. Therefore, a fair appraised grade is expressed as, $T = \backslash\{t_k, t \in \boldsymbol{G}$ and $t_{k-1} \geq t_k \leq t_{k+1}$ where k = index of true grade in $\boldsymbol{G}\backslash\}$. The benchmark scale in Fig. 12 is thereby rigorously followed so that the fair and unfavorable cases can be determined. Lower gainer and loser ratios are better as that would reduce the adverse effect among the youth. Maximum fairness and the lowest loser ratio are what make an appraisal model more credible to policy-makers. In Fig. 13, consistently all the seven subjects have a significantly high ratio in Fair-Zone which indicates around 82% of the time the model appraisal falls into tight fairness bound. If each course grade has the lowest deviation with the original grade point then an interrogator of the Turing test will not be able to distinguish an appraised transcript from originally obtained ones. Furthermore, while making a trade-off, the model is only unfavorable in each of the subjects, with an average loser ratio of 0.031. For an ideal appraisal model, the following are desirable in the litmus test:

1. Fairness close to 1.0 and Loser close to 0;
2. Optimal trade off with Gainer ratio such that point 1. is achieved;

The outcome from the automated appraisal system passes through

two iterations of possible loser flag correction. These extra iterations of flag prediction ensure that erroneous appraisal does not harm the candidates. If the supervised flag prediction model detects an unfair appraisal, then the candidate is given a higher grade if there is a higher grade with the second-highest probability. However, each grade prediction from the deep appraisal model is interpretable with the SHAP method. Using the Shapley additive explanation algorithm, the contribution of every feature to a prediction probability is interpreted.

Using the deep learning model explainer process, the deciding factors for output grade probability are shown as a positive and negative force in Fig. 14. The example shows the output probability of getting an A+ in a particular subject. The Explainer process, i.e., the SHAP method, determines an expected probability output $E[f(z)]$ for each output node, which is set as a base value in the force plot (Lundberg & Lee, 2017, Fig. 1). $E[f(z)]$ is the output if no features of the current output $\widehat{y_{c,i}}$ was known. In this example, the deciding factors for an A+ pushed the probability slightly above the base value. Input contributions for all letter grades (A+, A, A-, B, C & D) can be produced in parallel. For simplicity, only the top three features pushing the probability higher or lower than the base value are kept in Fig. 14 whilst the original interactive plot includes more than 400 attributes.

To evaluate whether the converged model could pass the Turing test mentioned in 3.1, it is indispensable to produce an indistinguishable distribution of results for each subject. Closer KL divergence to zero ensures that the two distributions are as similar as peas in a pod. Consistently, the appraisal model in this research showed lower KL divergence for each subject given in Table 3. The fairness ratio and KL divergence between the real outcome and model prediction indicate that the implemented model has achieved a reasonable performance to rationally appraise high-stakes assessment.

4.3. Data quality assessment

To discover the presence of a signal, a random dataset with the same dimension as the original training data was produced by performing a shuffle of the input vectors with the aim to train the appraisal model. If input matrix x has no signal, then the model would output the most observed grade during the training procedure. How the new appraisal model performed on real-world test data was determined from the confusion matrices for the obligatory courses in Fig. 15. The appraisal on test data revealed that the model that was trained on a random dataset could not learn from input signals as the confusion matrices for the compulsory subjects produced in Fig. 15 showed the appraisal model assigned maximum probability for the most encountered grade point. For example, as the letter grade distribution in Fig. 10, **A**-appeared most of the time as the target label for the Bangla subject. Therefore, the new model naively predicted an **A**-grade for Bangla as shown in Fig. 15 to minimize the categorical cross-entropy loss. Consequently, the erroneous appraisal of the Bangla course would allow a substantial number of graduates to become gainers (who would attain lower than **A-**) and losers (who would attain higher than **A-**). Likewise, for the other two compulsory subjects, English and ICT, the new model awarded the highest observed letter grades. Thus, the appraisal model that learnt to predict the HSC grades from the synthesized training dataset failed to rationally appraise the three compulsory subjects-: Bangla, English. and ICT. In contrast, the appraisal model linked with the original training data did not maximize its soft-max probability only based on the highest appearing grade, as illustrated in 4.2; therefore, the model that was trained with real-world historical data could appraise the exam grades from the pattern discovery between grades and features.

Table 3
fairness and kl divergence in each subject.

Fair Appraisal	KL Divergence
81%	0.19
80%	0.09
86%	0.44
78%	0.21
82%	0.46
86%	0.22
80%	0.12

Table 4
Deviation of average grade point distribution of two methods using KL Divergence.

	Machine Learning Appraisal	Weighted Average (baseline model)
KL Divergence	0.35	1.19

To ensure the quality of the training data, this study followed the iterative procedure (Angluin & Laird, 1988, Table 1)) to determine the noise rate η_b and compared the required number of training samples m for obtaining an appraisal model with a high probability that the appraised grades are not too different from the true outcome (Angluin & Laird, 1988, Theorem(2)). Firstly, the noise determination procedure outputs the rate of disagreement, i.e., the proportion of disagreement with θ during the initial training process of the appraisal model as an estimator for the noise η_b. The noisy examples were identified using the confidence joint (Northcutt et al., 2021, Fig. 1). The iterative search procedure converged on a small fraction of $\eta_b = 0.375$. Then, with the quantification of noise rate in the training sample, the lower bound of the required number of observations was determined using the equation in (Angluin & Laird, 1988, eq. (1)) in Equation (15). The right-hand side of the inequality in Equation (15) gave the lower bound for the required size of training sample m to insure the probability of correct appraisal of at least 80% (tolerance $\varepsilon = 0.2$) with a tiny probability of $\delta = 0.02$ that the training sample is wildly unrepresentative. The parameter N is the number of times the parameter space of the appraisal model is updated, i.e., the finite set of appraisal rules. The value of N is a large number, but it has a small contribution to the sample size due to the log operation.

$$m \geq \frac{2}{\varepsilon^2(1-2\eta_b)^2}\ln\left(\frac{2N}{\delta}\right) \tag{15}$$

The training samples were approximately four-fold bigger than the minimum bound in Equation (15). If Equation (15) did not satisfy, a new η_b was generated with the noisy instances discarded. In the worst case, the trade-off had to be made by increasing the tolerance level ε.

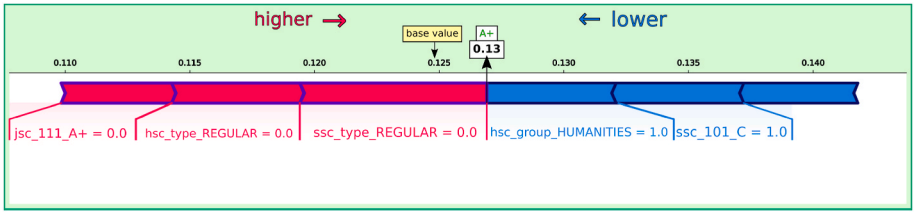


Fig. 14. Major forces by portfolio features on a candidate's grade appraisal.

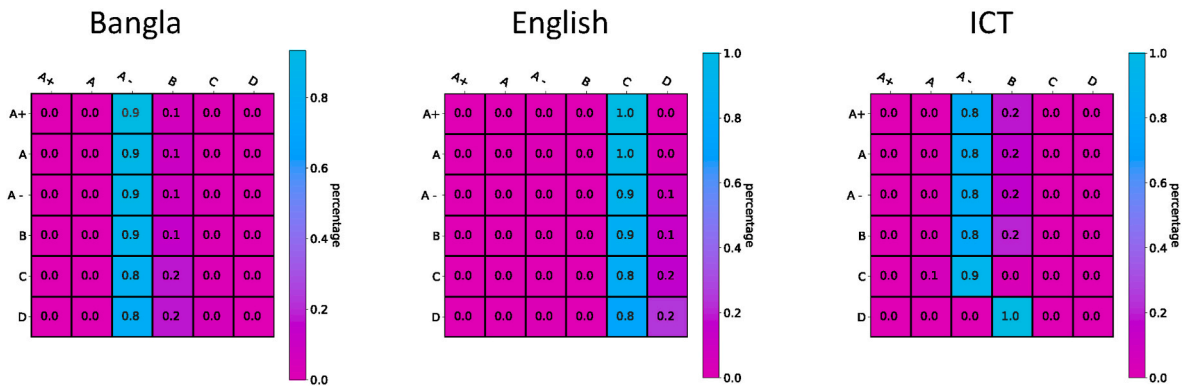


Fig. 15. Confusion matrices produced on test data performed by the model with synthesized training data.

5. Discussions

In response to Covid-19 in the year 2021, some of the E−9 member countries followed the weighted average method to appraise 12th grade students without the exam. The baseline model considers only two parameters: associated grade point of the junior school certificate and the secondary school certificate. However, the method is not inclusive in the sense that the process is not applicable to candidates who are irregular or have migrated to different study groups. The perceived fairness was one pitfall of the baseline model. For instance, if the compulsory subject 1 is appraised with the weighted average method where $w_{ssc\_bangla} = 0.75$ and $w_{jsc\_bangla} = 0.25$ then from this appraisal process, ≥20.12% of the candidates get grades higher than the +1 deviation (at least 20.12% gainer). Moreover, ≥3.27% of candidates become losers in the first compulsory subject (Bangla) if appraised by a weighted average method. While in compulsory subject 2 (English), ≥40% of students are given grades higher than +1 grade point, which makes the distribution and appraised grades substantially dissimilar to those in original transcripts. The unbounded gainer and lower proportion in each subject increase the risk of deviating a lot from the true distribution. Another disadvantage of the baseline model is the ceiling for HSC grade points. To illustrate, a student who obtained 3.5 in SSC and 3.5 in JSC, achieves 3.5 in the HSC, which is the maximum obtainable (3.5*0.75 + 3.5*0.25) from the weighted average model. But it is natural for a student with a 3.5 in previous board exams to obtain an A or A+ later in the HSC exam. Therefore, setting a ceiling for higher secondary results violates the stochastic nature of the real world. This section will focus on the national level distribution of the cumulative grade point average generated from the baseline model and the proposed model to the ground truth distribution. Deviation or dissimilarity from the CGPA distribution in the passing year of 2019 is analyzed using the KL Divergence metric. In Table 4, the proposed model happens to produce a closer distribution curve to the real world curve as the model learned to capture complex patterns in its latent space. On the other hand, the baseline model that considers only JSC and SSC grade points generated a distribution curve that has a KL Divergence of 1.19 with the real world curve. From the comparison shown in Table 4, a lower relative entropy, or KL Divergence, is better matched with the reference distribution, the true distribution of average cumulative grade points.

Therefore, appraisal through proposed machine learning intimates the real-world distribution better than the baseline model, which will prevent an anomalous distribution of HSC examination results. The proposed model, therefore, retains transparency, accountability, and fairness at the individual level as well as at the aggregated level. If any national education board wants to implement an alternative appraisal system, this study can help in benchmarking their solutions with the proposed model. Moreover, policy-makers and people will be informed about the level of fairness they can expect from the proposed appraisal model, which will eventually decrease policy-makers' skepticism and people's mistrust or fear of AI-driven transcript generators. As a thought experiment, consider the worst-case scenario wherein the lock-down is going to be extended for another couple of months, or another outbreak of locust attack, earthquake, or cyclone arrives. Will education policy experts be able to equitably protect the nation's future workforce and academic loss without creating an adverse effect? To achieve a unified and responsible AI appraisal system, this article presents a framework addressing the critical task of alternatively appraising high-stakes exam candidates during an emergency, guaranteeing that appraisal of candidates from quantitative and qualitative features is attainable with high fairness and minimal risk factor. From the findings shown in this study, it is important to know the reliability of AI-driven assessment tools in the future.As with the majority of studies, the design of the current study is subject to limitations. A major limitation in applying the methodology could be the inconsistency of the historical data. A recent policy change on the assessment will possibly make the previous portfolio information obsolete, resulting in obtaining fewer training data points. The limitation to adapting the proposed model is that, as opposed to an established assessment system, there are not enough institutional structures and policies for an alternative emerging model (Clarke, 2011). The authors perceive that the limitation of not being accurate 100% of the time is because this is the maximum achievable accuracy with the architecture. Therefore, the future direction will include meta-learning algorithms to dynamically generate the architecture of the appraisal model. Moreover, the future endeavor is to create a knowledge base from the trained model's latent variables so that policy-makers can identify factors that are responsible for higher secondary exam performance and improve them.

6. Conclusion

The study sets out an AI agent receiving maximum fairness and the lowest detrimental effect as an appraisal system for high-stakes exams. Any national emergency creates a barrier to arranging high-stakes exams in large education systems, such as the E−9 countries, causing candidates' progress to halt. In order to ensure fairness and equity to the entire batch of candidates, appraising subject-wise grades using a responsible AI model having strong predictive skills can be a rationale policy as an alternative to examination. This research suggests explainable AI as an equitable substitute for high-stakes exams during crisis situations. The rigorous empirical research scrutinizes the strength of the proposed inclusive computerized system, which was developed to appraise transcripts of a higher secondary standardized exam. The universal function approximation technique lies at the core of the appraisal model. The appraisal model will be useful even when there is no need to assess candidates with machine learning; predicting high-stakes exam performance beforehand, and ensuring extra care can be one use case. Not only will a rational machine learning model come in handy when policy-makers look for an alternative of the exam arrangement on a

large scale, but this will also build trust in AI and data science responses to epidemics that can mitigate potential harm. The appraisal framework provides high flexibility for policy-makers in terms of choosing output types, portfolio features, loss functions, activation functions, and other hyper-parameters associated with the framework. In a particular case, if the authorities decide to take exams on a reduced number of subjects and appraise the entire transcript on that basis, the proposed method can leverage the fairness of appraisal by generating the rest of the grades considering all the existing information as input. This trained model can further be used as a checkpoint for transfer learning.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of aic and bic. *Ecology, 95*, 631–636. URL: http://www.jstor.org/stable/43495189.

Aldrich, J. (1997). R. a. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science, 12*, 162–176. URL: http://www.jstor.org/stable/2246367.

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education, 17*, 1–21. https://doi.org/10.1186/s41239-020-0177-7

Angluin, D., & Laird, P. (1988). Learning from noisy examples. *Machine Learning, 2*, 343–370. https://doi.org/10.1023/A:1022873112823

URL:. In Bank, W. (Ed.), *The COVID-19 crisis response*. World Bank. https://doi.org/10.1596/34571. arXiv:https://elibrary.worldbank.org/doi/pdf/10.1596/34571 https://elibrary.worldbank.org/doi/abs/10.1596/34571.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*, 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Bonney, G. E. (1987). *Logistic regression for dependent binary observations* (pp. 951–973). Biometrics. https://doi.org/10.2307/2531548. URL:.

Bosch, N. (2021). Identifying supportive student factors for mindset interventions: A two-model machine learning approach. *Computers & Education, 167*, Article 104190. URL: https://www.sciencedirect.com/science/article/pii/S0360131521000671 https://doi.org/10.1016/j.compedu.2021.104190.

Brevard, P. B., & Ricketts, C. D. (1996). Residence of college students affects dietary intake, physical activity, and serum lipid levels. *Journal of the American Dietetic Association, 96*, 35–38. https://doi.org/10.1016/S0002-8223(96)00011-9. URL:.

Chen, K. Z., & Li, S. C. (2021). Sequential, typological, and academic dynamics of self-regulated learners: Learning analytics of an undergraduate chemistry online course. *Computers & Education: Artificial Intelligence, 2*, Article 100024. URL: https://www.sciencedirect.com/science/article/pii/S2666920X21000187 https://doi.org/10.1016/j.caeai.2021.100024.

Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020a). Application and theory gaps during the rise of artificial intelligence in education. *Computers & Education: Artificial Intelligence, 1*, Article 100002. URL: https://www.sciencedirect.com/science/article/pii/S2666920X20300023 https://doi.org/10.1016/j.caeai.2020.100002.

Chen, X., Zou, D., Cheng, G., & Xie, H. (2020b). Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of computers & education. *Computers & Education, 151*, Article 103855. URL: https://www.sciencedirect.com/science/article/pii/S0360131520300555 https://doi.org/10.1016/j.compedu.2020.103855.

Clarke, M. (2011). *Framework for building an effective student assessment system: Read/saber working paper*. World Bank. URL: https://eric.ed.gov/?id=ED553178.

Davidson, C. N., & Katopodis, C. (2020). In a pandemic, everyone gets an asterisk. https://www.insidehighered.com/views/2020/03/23/during-covid-19-crisis-higher-ed-should-rethink-how-assess-test-and-grade-students.

Dewan, M. A. A., Granger, E., Marcialis, G. L., Sabourin, R., & Roli, F. (2016). Adaptive appearance model tracking for still-to-video face recognition. *Pattern Recognition, 49*, 129–151. URL: https://www.sciencedirect.com/science/article/pii/S0031320315002903 https://doi.org/10.1016/j.patcog.2015.08.002.

Dewan, M., Murshed, M., & Lin, F. (2019). Engagement detection in online learning: A review. *Smart Learning Environments, 6*, 1–20. https://doi.org/10.1186/s40561-018-0080-z

Dhanalakshmi, R., Anuja Mary, A., Shrijith, D., & Vijayaraghavan, N. (2021). A study on covid-19 – impacting indian education. *Materials Today Proceedings*. https://doi.org/10.1016/j.matpr.2021.02.786. In press, URL https://www.sciencedirect.com/science/article/pii/S2214785321019465.

Educationboard Bd. (2020). Rules of business. http://www.educationboard.gov.bd/dhaka/rules_business.php.

Friedler, S. A., Tan, Y. L., Peer, N. J., & Shneiderman, B. (2008). Enabling teachers to explore grade patterns to identify individual needs and promote fairer student assessment. *Computers & Education, 51*, 1467–1485. URL: https://www.sciencedirect.com/science/article/pii/S0360131508000353 https://doi.org/10.1016/j.compedu.2008.01.005.

Guterres, A. (2020). Policy brief:education during covid-19 and beyond. https://unesdoc.unesco.org/ark:/48223/pf0000373247/PDF/373247eng.pdf.multi.

Hashan, M. J. (2020). Op-ed: When you are an hsc student in 2020. https://archive.dhakatribune.com/opinion/op-ed/2020/10/10/op-ed-when-you-are-an-hsc-student-in-2020.

Hassan, M. M., Alam, M. G. R., Uddin, M. Z., Huda, S., Almogren, A., & Fortino, G. (2019). Human emotion recognition using deep belief network architecture. *Information Fusion, 51*, 10–18. https://doi.org/10.1016/j.inffus. 2018.10.009

Helberger, N., Araujo, T., & de Vreese, C. H. (2020). Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Computer Law & Security Report, 39*, Article 105456. https://doi.org/10.1016/j.clsr.2020.105456

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*, 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Hossain, R. I. (2020). Hsc 2020 cancelled: You can't make everyone happy. https://archive.dhakatribune.com/bangladesh/education/2020/10/07/hsc-exams-cancelled-will-the-batch-of-2020-suffer-because-of-it.

Ichino, A., & Winter-Ebmer, R. (2004). The long-run educational cost of world war ii. *Journal of Labor Economics, 22*, 57–87. https://doi.org/10.1086/380403

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. CoRR abs/1412.6980. URL: https://arxiv.org/pdf/1412.6980.pdf.

Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education, 5*. https://doi.org/10.3389/feduc.2020.572367. URL: https://www.frontiersin.org/article/10.3389/feduc.2020.572367.

Larochelle, H., & Bengio, Y. (2008). Classification using discriminative restricted Boltzmann machines. In *Proceedings of the 25th international conference on Machine learning* (pp. 536–543). https://doi.org/10.1145/1390156. 1390224

Lemay, D. J., Baek, C., & Doleck, T. (2021). Comparison of learning analytics and educational data mining: A topic modeling approach. *Computers & Education: Artificial Intelligence, 2*, Article 100016. URL: https://www.sciencedirect.com/science/article/pii/S2666920X21000102 https://doi.org/10.1016/j.caeai.2021.100016.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist, 54*, 421–431. URL: http://www.jstor.org/stable/27836590.

Libralon, G. L., de Leon Ferreira, A. C. P., & Lorena, A. C. (2009). Pre-processing for noise detection in gene expression classification data. *Journal of the Brazilian Computer Society, 15*, 3–11. https://doi.org/10.1007/BF03192573

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777). URL: https://dl.acm.org/doi/pdf/10.5555/3295222.3295230.

Mamun, S. (2020). Hsc exams cancelled: Will the batch of 2020 suffer because of it?. https://www.thedailystar.net/shout/news/hsc-2020-cancelled-you-cant-make-everyone-happy-1974085.

Marinoni, G., van't Land, H., & Jensen, T. (2020). The impact of covid-19 on higher education around the world. *IAU Global Survey Report*. URL: https://www.iau-aiu.net/IMG/pdf/iau_covid19_and_he_survey_report_final_may_2020.pdf.

Mnih, V., Larochelle, H., & Hinton, G. E. (2011). Conditional restricted Boltzmann machines for structured output prediction. In *Proceedings of the twenty-seventh conference on uncertainty in artificial intelligence* (pp. 514–522). Arlington, Virginia, USA: AUAI Press.

Musso, M. F., Cascallar, E. C., Bostani, N., & Crawford, M. (2020). Identifying reliable predictors of educational outcomes through machine-learning predictive modeling. *Frontiers in Education, 5*. https://doi.org/10.3389/feduc.2020.00104. URL: https://www.frontiersin.org/article/10.3389/feduc.2020.00104.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47*, 90–100. https://doi.org/10.1016/S0022-2496(02)00028-7

News18. (2021a). Boy kills himself over poor marks in class 10 board exams. https://www.news18.com/news/education-career/boy-kills-himself-over-poor-marks-in-class-10-board-exams-4048673.html.

News18. (2021b). Cbse 10th result 2021: Delhi hc to hear plea seeking modification in assessment formula. https://www.news18.com/news/education-career/delhi-hc-to-hear-plea-seeking-modification-in-cbse-10th-result-2021-formula-cbse-nic-in-3943145.html.

News18. (2021c). Parents, students flag concerns over cbse, cisce result calculation formula. https://www.news18.com/news/education-career/parents-students-flag-concerns-over-cbse-cisce-result-calculation-formula-3874034.html.

Niko Kommenda, F. H. J. (2020). Covid vaccine tracker: When will a coronavirus vaccine be ready? *The Guardian*. URL: https://www.theguardian.com/world/ng-interactive/2020/oct/29/covid-vaccine-tracker-when-will-a-coronavirus-vaccine-be-ready.

Northcutt, C. G., Jiang, L., & Chuang, I. L. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research, 70*, 1373–1411. https://doi.org/10.1613/jair.1.12125

Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning. arXiv preprint arXiv:1811.03378 doi:https://doi.org/10.48550/arXiv. 1811.03378 .

OBeirne, M. D., Himes, M. D., Soboczenski, F., Zorzan, S., Baydin, A. G., Cobb, A. D., … Domagal-Goldman, S. (2019). Inara: Intelligent exoplanet atmospheric retrieval a machine learning retrieval framework with a data set of 3 million simulated exoplanet atmospheric spectra. In *2019 astrobiology science conference*. AGU. URL: https://agu.confex.com/agu/abscicon19/meetingapp.cgi/Paper/481266.

Pedro, A. J., Hasan, A., Goldemberg, D., Geven, K., & Aroob, I. S. (2021). Simulating the potential impacts of covid-19 school closures on schooling and learning outcomes: A

set of global estimates. *The World Bank Research Observer, 36*, 1–40. https://doi.org/10.1093/wbro/lkab003

Rahman, M., Hamzah, M. I. M., Meerah, T., & Rahman, M. (2010). Historical development of secondary education in Bangladesh: Colonial period to 21st century. *International Education Studies, 3*, 114–125. URL: https://eric.ed.gov/?id=EJ1066070.

Richard, M. D., & Lippmann, R. P. (1991). Neural network classifiers estimate bayesian a posteriori probabilities. *Neural Computation, 3*, 461–483. https://doi.org/10.1162/neco.1991.3.4.461

Rizvi, S., Rienties, B., & Khoja, S. A. (2019). The role of demographics in online learning; a decision tree based approach. *Computers & Education, 137*, 32–47. URL: https://www.sciencedirect.com/science/article/pii/S0360131519300818 https://doi.org/10.1016/j.compedu.2019.04.001.

Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., & Cascallar, E. (2021a). Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. *Computers & Education: Artificial Intelligence, 2*, Article 100018. URL: https://www.sciencedirect.com/science/article/pii/S2666920X21000126 https://doi.org/10.1016/j.caeai.2021.100018.

Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., & Cascallar, E. (2021b). Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. *Computers & Education: Artificial Intelligence, 2*, Article 100018. URL: https://www.sciencedirect.com/science/article/pii/S2666920X21000126 https://doi.org/10.1016/j.caeai.2021.100018.

Saddiqa, M., Magnussen, R., Larsen, B., & Pedersen, J. M. (2021). Open data interface (odi) for secondary school education. *Computers & Education, 174*, Article 104294. URL: https://www.sciencedirect.com/science/article/pii/S0360131521001718 https://doi.org/10.1016/j.compedu.2021.104294.

Salakhutdinov, R., Mnih, A., & Hinton, G. (2007). Restricted Boltzmann machines for collaborative filtering. In *ICML '07: Proceedings of the 24th international conference on Machine learning* (pp. 791–798). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1273496.1273596. URL:.

SDG-Education 2030 Steering Committee. (2020). *The sdg-education 2030 steering committee recommendations for covid-19 education response*. https://sdg4education2030.org/sdg-education-2030-steering-committee-resources.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature, 529*, 484–489.

Solano-Flores, G., Raymond, B., Schneider, S. A., & Timms, M. (1999). Management of scoring sessions in alternative assessment: The computer-assisted scoring approach. *Computers & Education, 33*, 47–63. URL: https://www.sciencedirect.com/science/article/pii/S0360131599000184 https://doi.org/10.1016/S0360-1315(99)00018-4.

Szenicer, A., Fouhey, D. F., Munoz-Jaramillo, A., Wright, P. J., Thomas, R., Galvez, R., Jin, M., & Cheung, M. C. M. (2019). A deep learning virtual instrument for monitoring extreme uv solar spectral irradiance. *Science Advances, 5*. https://doi.org/10.1126/sciadv.aaw6548. arXiv:https://advances.sciencemag.org/content/5/10/eaaw6548.full.pdf.

The Indian Express. (2021). Explained: What is cbse's formula for evaluating class xii students' results?. https://indianexpress.com/article/explained/cbse-class-12-board-exam-result-evaluation-method-explained-7363291/.

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/tnnls.2020.3027314

Turing, A. M. (1950). Computing machinery and intelligence. Mind LIX. URL: https://doi.org/10.1093/mind/LIX.236.433.

UNESCO's COVID-19 Education Response. (2020). Managing high-stakes exams and assessments during the covid-19 pandemic. https://unesdoc.unesco.org/ark:/48223/pf0000373247/PDF/373247eng.pdf.multi.

Wan, X. (2019). Influence of feature scaling on convergence of gradient iterative algorithm. *Journal of Physics: Conference Series, 1213*, Article 032021. https://doi.org/10.1088/1742-6596/1213/3/032021. URL:.

Wang, Y. (2005). A multinomial logistic regression modeling approach for anomaly intrusion detection. *Computers & Security, 24*, 662–674. URL: https://www.sciencedirect.com/science/article/pii/S0167404805000751 https://doi.org/10.1016/j.cose.2005.05.003.

Wang, J., Hwang, G. H., & Chang, C. Y. (2021). Directions of the 100 most cited chatbot-related human behavior research: A review of academic publications. *Computers & Education: Artificial Intelligence, 2*, Article 100023. https://doi.org/10.1016/j.caeai.2021.100023

Xinhua. (2020). Bangladesh cancels major public examination amid covid-19 fears. http://www.xinhuanet.com/english/2020-10/07/c_139424338.htm.

Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers & Education: Artificial Intelligence, 2*, Article 100008. URL: https://www.sciencedirect.com/science/article/pii/S2666920X21000023 https://doi.org/10.1016/j.caeai.2021.100008.

Zhang, K., & Aslan, A. B. (2021). Ai technologies for education: Recent research & future directions. *Computers & Education: Artificial Intelligence, 2*, Article 100025. URL: https://www.sciencedirect.com/science/article/pii/S2666920X21000199 https://doi.org/10.1016/j.caeai.2021.100025.

Zhu, X., & Wu, X. (2004). Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review, 22*, 177–210. https://doi.org/10.1007/s10462-004-0751-8