education sciences

MDPI

Article

Technology-Enhanced Learning, Data Sharing, and Machine Learning Challenges in South African Education

Herkulaas MvE Combrink *, Vukosi Marivate and Baphumelele Masikisiki

Department of Computer Science, University of Pretoria, Pretoria 0028, South Africa
* Correspondence: u29191051@tuks.co.za

Abstract: The objective of this paper was to scope the challenges associated with data-sharing governance for machine learning applications in education research (MLER) within the South African context. Machine learning applications have the potential to assist student success and identify areas where students require additional support. However, the implementation of these applications depends on the availability of quality data. This paper highlights the challenges in data-sharing policies across institutions and organisations that make it difficult to standardise data-sharing practices for MLER. This poses a challenge for South African researchers in the MLER space who wish to advance and innovate. The paper proposes viewpoints that policymakers must consider to overcome these challenges of data-sharing practices, ultimately allowing South African researchers to leverage the benefits of machine learning applications in education effectively. By addressing these challenges, South African institutions and organisations can improve educational outcomes and work toward the goal of inclusive and equitable education.

Keywords: data-sharing governance; machine learning education research; challenges; innovation; South African context

Citation: Combrink, H.M.; Marivate, V.; Masikisiki, B. Technology-Enhanced Learning, Data Sharing, and Machine Learning Challenges in South African Education. *Educ. Sci.* **2023**, *13*, 438. https://doi.org/10.3390/educsci13050438

Academic Editor: Peter Williams

Received: 14 February 2023
Revised: 18 March 2023
Accepted: 5 April 2023
Published: 24 April 2023

1. Introduction

Machine-learning-based research relies heavily on the availability of data. In the event that there is no access to relevant data or there is a small dataset, this type of research can be delayed or produce poorly performing and unreliable models due to the lack of relevant data. The importance of data sharing for machine-learning-based research has led to an acceleration in machine-learning implementation and higher accuracy [1].

Unfortunately, data access is not always a straightforward process, because every country and the academic institutions within these countries are governed by policies and regulations that are implemented differently. For example, in Europe, the General Data Protection Regulation (GDPR) has been questioned by a number of researchers, especially in the biomedical domain [2]. They have stipulated that consent for data and bio-sample sharing generates confusion and uncertainty and creates conflicts between the regulations and research ethics [2–5]. This is because, for biological data, understanding the context of each biological entity; their circumstances; and (in the case of human participants) their demography, gender, lifestyle, diet, and general behaviour is important [4]. As a result, some researchers are reluctant to share data because of legal fears and social sanctions in combination with the threat of huge penalties as a consequence of violating the GDPR [6]. In South Africa, protection legislations such as the Protection of Personal Information Act (POPIA) and ethical policies are in place [7]. POPIA refers to the protection of distinct personal information, which includes information on race, ethnic origin, political persuasions, sex life, and other factors. However, concerns have been raised by various researchers as to whether POPIA conflicts with the existing data protections and research ethics [8]. Thaldar and Townsend (2021) argued that POPIA is a positive development for data privacy; however, it places certain restrictions on the distribution and sharing of data,

which impacts the way data are accessed and shared. This argument further extends to the idea that through these policies, the practice of data sharing might be compromised and not the act of generating data that is deidentified in the correct manner [5]. The stringent regulations, although cumbersome, are intended to protect personal information and members of the public from the potential exploitation of their shared information [9]. The multitude of risks associated with personal information ending up in the wrong hands cannot be overstated, as there are legitimate and serious consequences to the sharing of personal information [10]. We focus on how some of these policies and regulations impose restrictions on the access and sharing of data. In our argument, we pose that these restrictions result in delaying academic research—especially the implementation of machine learning for educational research (MLER)—and other domains that require large quantities of training data to build innovative technologies. Longo and Drazen (2016) defined data sharing as the disposition and preservation of data for public access with the purpose of providing access for reuse [11]. Chawinga and Zinn (2019) defined data sharing as the publication of the primary data and any supporting materials required to interpret the data acquired as part of a research study [12]. Dietrich et al. (2014) argued that researchers should not practice data sharing among themselves without considering the legal and ethical procedures, and they further defined data sharing as a practice of making data available and accessible to other academics, taking into account all the legal and ethical implications associated with its sharing [5].

The Promotion of Access to Information Act requires the Department of Education in South Africa to make education-based data publicly available to academics in education and ensure that the confidentiality of data is not breached [6]. As simple as this sounds, the frequency of data being shared, as well as the type of data, is not conducive for the MLER research needed to implement 4IR technology in the education domain [13]. As it currently stands, the data shared are too aggregated and cannot be implemented for machine learning use, as can be seen in the Department of Higher Education and Training Annual Reports for various institutions [14]. The current data presented in these reports are not the kinds of data we are referring to—instead, we make claims relating to data such as student gradebook information, assessments of students collected longitudinally, and laboratory reports. We agree with the perspective that the current course of technological growth in MLER for the African continent, and, moreover, South Africa, is at risk of taking a different trajectory favouring South Africans as consumers rather than innovators of such technology because of data sharing and how it is governed [15]. This unfortunate potential outcome might become a reality if data-sharing practices, collaboration, and a much-needed skills revival within the basic education system do not take place [16]. To avoid this possible scenario, it is important that the complexity of the situation is outlined, drawing from a variety of different factors including how South African innovation is currently viewed in the education research space, how MLER can be advanced, and how data protection might stifle technological innovation and advancement if it is not managed properly [17]. Therefore, the purpose of our study was to scope the current landscape through the context of MLER in South Africa and identify key concerns that need to be addressed for the advancement of MLER in this context.

2. Research Design and Method

2.1. Research Design and Methodology

To gauge the landscape of data sharing for the MLER context, a scoping review was used as an assessment methodology [18,19]. The scoping review was performed to assess the current information related to data sharing for MLER within the South African context. The scoping review aimed to provide a conceptual overview of the data-sharing practices present within institutions of higher learning, highlighting the challenges and current practices associated with data sharing. The purpose of a scoping review is to provide the clarity required within a specific field of study from the perspective of addressing a broad research question and aim conceptually, within a specific domain that requires further

contextualisation [20]. This is performed by assessing secondary data sources, the literature, and expert opinions about a particular subject matter.

The overarching question that was explored was guided by a need to assess the data-sharing practices present within institutions of higher learning. This methodological approach was favoured as there are gaps in the understanding related to data-sharing practices that may have obstructing impacts on MLER research. The interpretation of the results was discussed from an interpretivist paradigm. This was proposed due to the subjective nature of the information, as the bodies of knowledge included different domain-specific contexts. In order to minimise the potential bias in the interpretation of the experts, the themes identified were used as the discussion points, and the arguments were developed based on the findings. This scoping review allowed for the themes to be identified from the literature available so that further discussion on the topics could commence.

2.2. Inclusion and Exclusion Criteria

The scoping data collection strategy included a specific scope and selection criteria for inclusion and exclusion. Initially, more than 80 relevant academic articles within the local (South African) and global contexts were considered within the scoping review. These articles contained keywords related to data sharing, machine learning for education, challenges in obtaining training data, data-sharing policies, and GDPR and POPIA with regard to data-sharing frameworks. Furthermore, only works obtained from 2010 onward were considered for the review (Table 1).

Table 1. Inclusion and exclusion criteria in the study.

Inclusion Criteria	Exclusion Criteria
Information after 2010. Keywords: data sharing, machine learning for education, challenges in obtaining training data, data sharing in education, data sharing policies, GDPR, POPIA, data sharing frameworks, South Africa, education.	No information prior to 2010. If the body of knowledge was only about a specific organisation, and not the data sharing between organisations, or if the data sharing involved criminal activity or examples that were against the law.

Exclusion criteria were applied once the literature was gathered. The exclusion criteria were applied if the body of knowledge only made reference to a very specific use case within an organisation (rather than the organisation itself), whether or not the principles could be applied to a context involving multiple institutions (for example, if a specific South African institution of higher learning had a policy that was only specific to that institution, then it was removed from the criteria); or if the data sharing had anything to do with criminal activity or the use of data for negative use cases (as this was not the scope or the focus of this very specific study). Upon the initial screening of the body of text, 50 items of literature were excluded, given the exclusion criteria. In addition to this, a bibliometric analysis was not included within the current study, as the interpretation of the body of the texts was the focus of the scoping review, rather than the statistics and counts of the number of literature items within this field of research. In total, 34 items of literature in the form of journal articles (28), conference proceedings (5), and a book review (1) were considered for the scoping review [21–53].

2.3. Analysis

The most salient ideas were summarised from each of the literature sources and coded according to their themes. The thematic coding of ideas was performed within the scoping process and noted on a separate data source. The thematic coding was conducted on the premise of the most salient ideas surrounding data sharing, governance, and the challenges associated with them. The analysis was only performed on the final data included in the scoping review. An initial analysis was performed to identify the

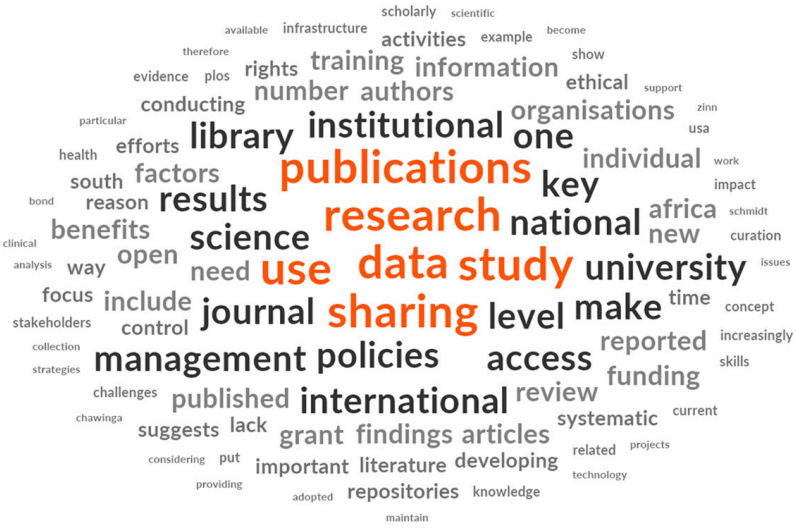
broad overarching categories, which were then subdivided to understand the nuance and gain further context for the subject matter. Each of the broadly defined categories underwent another thematic analysis to extract the themes within each of them. This meant that the analysis process occurred on two different levels of understanding for the texts: (a) a superficial categorisation of the broader themes from within the literature, and (b) a more descriptive and nuanced analysis of the text to understand the scope and context of the categories. The information was then further prepared for analysis using Python 3.7 for the data processing of the frequency graphs and the visualisation of the codified data. These visualisations included word clouds to illustrate the most salient synonyms and related words within the body of knowledge that were grouped into the respective themes. Furthermore, the information was processed to include additional descriptive data such as word frequencies and frequency graphs and tables. These datapoints were then grouped to include the primary overarching ideas related to data sharing in this context for further discussion. The summary of these results is outlined within the results section below.

3. Results

Based on the first iteration of coding the data, three primary categories were identified for the scoping review, namely: different definitions of data sharing; a need to upskill experts, share data, and protect personal information; and data sharing and the impact on research.

3.1. Different Definitions of Data Sharing and the Impact on Research

Based on the literature consulted in the scoping review, the content related to definitions contained the following keywords and themes (Figure 1).



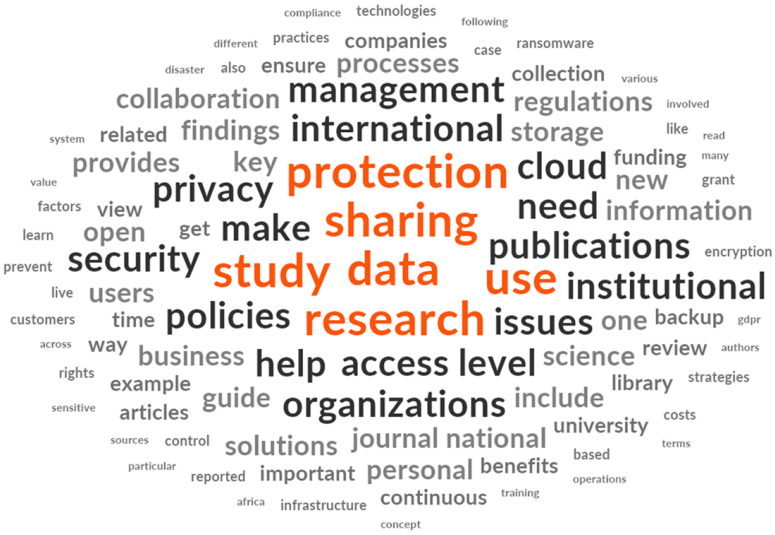
Name of Theme	Description
Availability of data	Data sharing needs to be made shareable to different stakeholders
Data use and reuse	Data can be used for multiple purposes and that is why sharing is vital
Data sharing has legal consequences	Data sharing should be performed with the correct frameworks in place
Data for research should be treated differently	Research data should advance the field and promote innovation

Figure 1. Frequency of words and themes used to define data sharing.

Data-sharing definitions differ between experts, and part of the scoping review was aimed at defining this concept. Data sharing across various domains and contexts has an impact on the type of practices that influence to what extent information is shared between different organisations in the context of MLER and other research-related activities involving the use of digital information. In the context of this scoping review, the definitions considered included a variety of keywords and concepts that overlapped and some that differed on the basis of the use case. For example, medical information sharing specific to patients differed from aggregate data about education. The four subthemes identified were the most prevalent across the literature sources and included a variety of concepts related to data reuse, legal consequences, research data as opposed to industry data, and the availability of data.

3.2. A Need to Upskill Experts, Data Sharing, and the Protection of Personal Information

Three salient themes within the expert literature obtained from the scoping review included a need to train and upskill experts in academia and industry, a need to share information more frequently between researchers, and a need to perform this sharing under very specific frameworks to protect personal information (Figure 2).



Name of Theme	Description
Upskill experts	Experts within academe, management, institutions, policy makers and those publishing should be data fluent
Share data	Data sharing should be universal between researchers and used to advance the field
Protect personal information	All data that are shared should protect the people whose data are used. The unintended consequences of such data shared should be considered

Figure 2. Frequency of words used to identify the themes related to the upskilling of experts, sharing data for research, and protecting personal information.

The themes identified are unpacked in the discussion section below. After the discussion of the themes, based on the scoping review, a call to action is presented, including the recommendations and proposed areas that require improvement for MLER to be successful in the South African context.

4. Discussion

Abebe et al. (2021) argued that machine-learning-related research is not only affected by data policies and regulations, but also a lack of knowledge about the value that data brings for these types of research and the insufficient technological resources that are used to transfer and share large datasets [21]. Another problem is related to the capacity of researchers and governing entities to assess not only the direct implications of the research itself, but also its unintended consequences, which is becoming increasingly important when machine learning and artificial intelligence (AI) are used within a system [34]. Enakrire (2020) maintains that in South Africa there is still a need for data literacy awareness programs that can aim at stimulating the growth of data-sharing practices in order to benefit South African research at a large scale [35]. Then, issues arise related to the availability of data that should be in the public domain [36]. Arnott et al. (2019) stated that the key obstacle to the success of machine learning is not the maturity of AI tools and techniques, but access to data in a sufficient volume and quality [37].

For academic research, organisations such as the Academy of Science of South Africa are promoting the need to increase data-sharing awareness, accessibility, and visibility for all academics [38]. The Academy of Science of South Africa is responsible for making sure that data are as open as possible for researchers [21]. However, some researchers have argued that it is unclear how ethics committees work in practice, and that some of these policies and regulations related to data access are unclear and bring complexity and unfairness to data sharing [4,39]. One of the challenges researchers face is related to the fundamental frameworks of which data can and cannot be shared and addresses pertinent questions such as ”how should these data be shared?”, ”where did the data come from?”, and ”what enabled the free sharing of this information?” [40–42]. In a variety of

contexts, new technologies (often imported from countries outside the African continent) are embedded with AI to make them function, which is more often than not trained on African data [43].

Industry giants with good financial standing can purchase data and build technologies without going through the governing gatekeeping mechanisms present within research institutions in third-world countries [44,45]. This is an indirect and unfortunate consequence of implementing governance structures that are too stringent within developing countries, as they enable an opportunity for industry to innovate in this space only if there is enough return on investment for the international counterpart—often only providing enough resources for the local (South African) adoption of the technology and not the capacity to innovate within the domain as well [46].

Abebe et al. (2021) argued that well-established data sharing could expand knowledge amongst researchers and contribute to better scientific practices [21]. In Africa, data-sharing practices can encourage collaboration between researchers and result in more research innovations [21]. Thaldar and Townsend (2021) echoed the notion that making data available to other researchers could elevate scientific research and create opportunities for more reproducible and accurate results [8].

Some data owners have their data shared and accessed, but do not reap the same benefits as the data collectors [14]. A study conducted by Riggs et al. (2019) called for global genomic data sharing [22]. In their study, they discussed the urgent need for more genetic data studies, because what they were trying to achieve did not have enough supporting data to innovate at the scale proposed [22]. Alter and Vardigan (2015) argued that some training pertaining to data should be available in a variety of formats and avenues to ensure the widest dispersion of knowledge [3]. Depending on the context of the research being conducted, the specifics of the dataset required vary; however, in most cases, the required machine learning data are always anonymised, and the identity of the research participants are not revealed [23]. For example, if the research intent is to predict a student's passing or failing of a course, the dataset that is usually required is the students' marks without their names; even if the data needed comprise more than just student marks, the Protection of Personal Information Act (POPIA) is always considered [20]. Furthermore, if the intent of the research is to automate assessment methods, the dataset that is needed is the answers and assigned scores, not the identity of students [24]. Data sharing is a crucial element of research and scientific advancement and a key to economic growth; it is therefore significant to have data shared fairly in all education disciplines [23]. Contrary to this, international organisations have free reign to explore the data, can perform any analysis they see fit, and can even sell the data, all under the banner of "improving the user experience" [25]. This scenario is applicable to not only learning management systems, but also any device or software that is internet-compatible [26]. A further complication in the current South African data-sharing ecosystem is the lack of collaboration between researchers within and between institutions [27]. As much as the strict data-sharing policies stifle innovation and growth in South African MLER, a drastic paradigm shift is needed to evaluate collaborative research [28]. The current Department of Higher Education and Training (DHET) funding model favours single-authored research over joint-authored research because the financial incentive to produce research per author decreases the more authors there are [5,22,29,30].

Dietrich et al. (2014) stated that researchers should not just share data among themselves without taking into account the legal and ethical considerations, and they defined data sharing as a practice of making research data available to others, taking into account the legal and ethical implications associated with their sharing [5]. However, Staunton et al. (2019) differed from Michener (2015) by pointing out that when following the legal process of data sharing, academic institutions might fail to differentiate between their handling of research data and institutional administrative data, and all of this ends up creating difficulties for data access [30]. They further argued that academic institutions in South Africa lack consistent and coherent policies and standards to govern data access and sharing [30]. In support of Staunton (2019), ethics review requirements and processes are

not straightforward. Firstly, there is a need to assess the strategic goals that various sectors must achieve, followed by the strategic goals of the specific institutions [31]. Thereafter, the specific policies that govern research activities and set targets for research goals need to be met [32]. To this extent, being able to conduct basic research or applied research without large datasets is becoming more difficult [33].

The movements in favour of data sharing and data transparency are gaining momentum globally. Some positive data-sharing practices have been published—for example, the work of the International Network for the Demographic Evaluation of Populations, which was carried out in places such as Africa, Asia, and Oceania by establishing a data repository in order to enable the sharing of fully documented and high-quality datasets [3].

In the South African context, MLER is not yet an implicit part of daily life, and few people realise the extent to which AI is developed, used, and governed in other contexts outside the African continent. One reason for this possible oversight can be seen in the rate at which machine learning and AI-compatible devices and the supporting technology were introduced into society [47]. For example, as the hardware of mobile devices became more and more sophisticated, so too did the software and the supporting machine learning elements of the devices [48]. South African basic education is filled with challenges in the digital domain, which has a spill-over effect into tertiary education. South Africa's basic education crisis is also fuelled by a deficit in basic education related to numeracy and literacy. According to McCann et al. (2021), the South African education crisis has had dire consequences, all related to the future of South Africa as a country and the livelihood of its people [49]. McCann et al. (2021) further argued that the education crisis has impacted innovation, skill transfer, and economic growth, which in turn has had an indirect impact on education, and so on [49]. Their arguments outlined fundamental ideas that sketched a scenario of society ultimately leading to entrapment and a lack of freedom, because with poorer education comes less and less economic freedom; less autonomy; less competence; less societal support; and, ultimately, a country with more societal challenges to be alleviated. For research to actively impact society, there needs to be not only an ecosystem that allows for research-related skills to exist within a specific society, but also mechanisms within that society to enable the acceptance of research and innovation. These mechanisms include the sharing of relevant data so that momentum can be added to the innovation in MLER. Unfortunately, the reality is that MLER is faced with significant unintentional challenges related to the sharing of data. The main reason for these data-sharing policies to be in place is to prevent personal identifiable information from leaking to anyone. This is important as there are different levels of identifiable information, but it is significant that a certain level of specificity and a certain level of aggregation are still needed to innovate in the MLER domain. It is possible to innovate if the variables used are categorised and aggregated, all identifiable information is removed, and only the target variables needed for prediction are used to drive the innovative technology. MLER is thus focussed on studying and using data science approaches to identify trends and anomalies in student data. This type of research does not (and should not) focus on identifying individuals based on who they are. A lack of understanding at a policy and governance level may provide insight into why a blanket approach is taken to data sharing, rather than a contextual, case-by-case approach.

Within the scope of MLER, predicting whether a student will pass or fail, or whether a student is under a state of cognitive duress and in need of medical assistance, is possible. However, as with any new technology and innovative solution, a series of testing and evaluation steps needs to be conducted before this can be implemented in real time. We therefore argue that data governance policies that are put in place to protect the personal information of individuals within an education setting can be revised to fully support MLER. This could be achieved through the promotion of data transparency awareness by machine learning researchers to reveal to the ethics committees that MLER poses no threat to the policies of data sharing. However, there are distinctions that need to be made between research to advance a field and the stringency required to implement technologies

that leverage machine learning. We believe that current policies overlook the impact that software technologies and computer learning software will have on education in South Africa, and perhaps the opportunities that MLER could bring in this domain and how automated decisions could be of use are not fully understood. Gatekeeping should therefore not occur at the research level, but rather at the stage of commercialisation, so that the unintended consequences may be unpacked from all perspectives. We also believe that there is a need to capacitate the current innovators and researchers across different fields so that critiques, academic inputs, innovation, and focus are all aligned based on the fields themselves and the impact that machine learning in education has on society. Unfortunately, this limits the scope of MLER and may generate new digital divides, all of which are related to a lack of public data and data sharing [30]. The biggest innovators of MLER at the moment are international industries, and not the academic institutions that are trying to conduct MLER. At the same time, and while South African researchers are trying to upskill and explore innovations in MLER, there are limits to the availability of data as a result of strict data-sharing policies. This vicious cycle further limits the extent to which South African researchers can innovate using their own data, while non-South African international companies building technologies for South Africa, using South African data, can explore and innovate using the same information. It must be mentioned that there are datasets available for the education domain to conduct MLER; however, the available datasets are not always contextually relevant. The extent to which relevant data need to be used for MLER is as vital as using the correct demographic and contextual features to conduct research in this domain. An example of a good dataset for MLER is the Open University (OU) virtual learning environment dataset (OULAD) [52]. The OU virtual learning environment dataset is an open-source, award-winning dataset intended for education research, including MLER research. This dataset contains information from approximately 32,000 students on social science and STEM courses and contains all the online click data accumulated from the OU virtual learning environment. Despite this incredible resource, there are convincing counterfactual arguments that nullify using this dataset in relation to African and, more specifically, South African MLER, because it is contextually different. What is lacking at the moment is the strategic alignment of fundamental goals for researchers in this domain to synergistically move toward technological innovation modelled for South Africa. Defining these terms should be driven by a shared vision if South African MLER is to remain relevant.

5. Call to Action

Based on the discussions, arguments, and outcomes from the scoping review, we propose an adapted data-sharing definition that includes MLER, as followings: "Data sharing for MLER is the practice of making research data available under different conditions, with the aim of promoting scientific advancement, social good, commercialisation, transparency, reproducibility, and/or collaboration. It involves sharing data and information under specified terms and conditions to ensure the responsible use and protection of research subjects' privacy as well as to take cognisance of the unintended consequences of sharing such information". MLER research may require context and data from African perspectives for each individual country and institution, and a lot of the data needed for MLER are not present within this network. Furthermore, even though there are some positive results from data-sharing practices between different countries, data-sharing practices are as yet not firmly established in all academic disciplines and countries regarding the training data needed for MLER. We therefore suggest that the education domain in South Africa has not yet fully established the practice of data sharing for MLER, because the data that are required for this kind of research necessitate expertise in identifying, storing, and sharing the information in a manner that complies with the current legal and ethical academic frameworks. We acknowledge the stringent frameworks that are currently in place, but we also suggest that there are some areas that require improvement.

In this study, we also echoed Riggs et al. (2019) by stating that there is an urgent need for fair data-sharing practices for MLER [22]. We argued that machine-learning-based research in education is not only impeded by data policies and regulations, but also by a lack of understanding on the part of regulatory bodies as to how MLER and data work together. We therefore fear that if these regulations remain at a hardware and data level, and do not include components of machine learning for education, they run the risk of stifling innovation, research and development, and transformation within the field because the data-sharing policies are too strict, focusing too much on the sharing of information rather than coding and aggregating information in ways that still facilitate education research.

With reference to Alter and Vardigan (2015), we outline some of the data types that are usually requested to conduct machine learning research in education [3]. We make this claim based on the lack of processes currently available to request and gain access to training data between different South African higher education institutions. For example, if a South African researcher wants access to data, extensive research proposals, ethical clearance, and a series of anonymisation steps are required just to obtain the data, but companies implementing learning management systems may conduct "in-house" research to "improve customer satisfaction" without such processes. Furthermore, researchers need to have a very specific research outcome in mind, a clear hypothesis, and a clear scope of analysis, leaving very little room to explore the data outside of the well-defined scope of the approved research proposal, thereby potentially stifling innovative research, whereas industry may have more freedom to explore the data. As stated above, this is the result of different policy implementation strategies managed differently between sectors and between countries internationally. International organisations and international companies with enough fluid capital are able to fast-track these processes under the guise of legal gatekeeping mechanisms, as the governance of policies (such as GDPR and POPIA) can be justified. This does not mean that the ethics committees themselves are redundant in the process of protecting information, but it does pose concerns if South African researchers require ethical clearance to conduct research from local committees, but international organisations do not when using the same data. We are not advocating MLER as the ultimate solution to the plethora of challenges that South African education faces, but as we outline, MLER enables highly innovative solutions in this context. For MLER to be successful, the underlying data-sharing governance structures need to support innovation in this domain. South Africa needs a shared strategy to achieve a more inclusive economy, taking account of how we share information between ourselves [19,23,29,50,51]. The current collaborative research model and the current ecosystem need to shift their paradigms to an objective, value-laden approach [53]. The set of values on a strategic level within the South African scientific community and general public (outside the sciences) should be shared between a variety of different stakeholders, and a more open network of researchers and research should be promoted in South Africa. We propose to achieve this by strategically involving all the relevant stakeholders to agree upon the values that the scientific community should focus on if we are to remain relevant in the MLER domain. For MLER-related research, we propose that the South African research community must uphold a set of common goals to strive for, as a collective, in order to shift the needle for research in this domain. Secondly, the epistemic value of science and innovation in South Africa should include a variety of different disciplines that are not traditionally seen as strictly within the domain of MLER research. This includes philosophy, general humanities, education, public relations, law, and commerce—all of which are important to drive this ecosystem. To define what the scientific community and research in these fields should prioritise, and where the value of research should lie, also requires an alignment of the current strategic goals. A variety of scientific goals and visions have been outlined by the 4IR commission in South Africa, the strategic vision of the Department of Science and Innovation (DSI), the Council for Scientific and Industrial Research (CSIR), and several other major South African institutions [47]. We propose that the education and ethical specialists both locally and internationally that work with South African data come together and collaboratively contribute toward advancing

the field of MLER research. Thus, we created a repository that can be accessed by any such expert with an internet connection, so that these meaningful contributions can be consolidated in a collaborative manner (https://github.com/dsfsi/Higher_Education_EDA accessed 21 January 2023). Within the repository, the proposed set of values currently outlined were not based on the collaborative efforts of the various stakeholders, but rather on the survey conducted to streamline these efforts and the arguments made in this paper from a practitioner perspective. It is therefore intended to serve only as a starting point for these shared dialogues and should be collaborative, as the core objectives might change depending on how the values are defined. It is also important that if we as a community of researchers strive for digital independence, we also address software ownership, copyright, and intellectual property based on the data we generate. Although these constraints might seem solely economically driven, the extent of innovation is dependent on the data and the skills needed to develop within those environments. There is also a need to specify which types of data and variables are acceptable to share and use for MLER, and which are not. For this to work, a consensus must be reached among researchers, policy makers, institutions, and industry so that MLER can advance. given the data-sharing governance and implementation structures. As outlined above, our arguments do not stem from the perspective of rebutting data-sharing policies, and we acknowledge that these policies are a vital part of the 21st century. Instead, our arguments centre around the rate of innovation and the type of innovation needed in developing countries, such as South Africa, that need innovation and technology to assist the education domain.

6. Conclusions

Data sharing, the current ecosystem, and the policy makers themselves have a significant hand in MLER. For the South African context, MLER is vital for advances in basic education. We propose that the objective value-laden approach should include how we involve the people whose data we use and at what stage of the research and commercialisation cycle we require the different policies at hand. This means that there needs to be a consideration of how researchers and companies involve these people when they produce an innovation, if the innovation is primed from their data. Data are a resource that fuels innovation, and without it, innovation becomes quite challenging. Science is objective, and scientific results are intended to be repeatable under the same conditions, serving as a foundation for future research. It is therefore recommended that when a researcher, either locally or internationally, conducts MLER in the context of South Africa, they should align with a set of values focussed on mitigating the unintended consequences that might arise as a result of innovation in this domain.

Author Contributions: Conceptualization, H.M.C., V.M. and B.M.; methodology, H.M.C.; software, H.M.C. and B.M.; validation, V.M. and B.M.; formal analysis, H.M.C., V.M. and B.M.; investigation, H.M.C., V.M. and B.M.; resources, V.M.; data curation, H.M.C.; writing—original draft preparation, H.M.C., V.M. and B.M.; writing—review and editing, V.M. and B.M.; visualization, H.M.C.; supervision, V.M.; project administration, V.M.; funding acquisition, V.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data generated in this study as well as additional insights may be found in the publicly shared higher education GitHub repository https://github.com/dsfsi/Higher_Education_EDA, accessed on 23 January 2023.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Messaoud, S.; Bradai, A.; Bukhari, S.H.R.; Quang, P.T.A.; Ahmed, O.B.; Atri, M. A survey on machine learning in internet of things: Algorithms, strategies, and applications. *Internet Things* **2020**, *12*, 100314. [CrossRef]
2. Morley, J.; Floridi, L.; Kinsey, L.; Elhalal, A. From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. In *Ethics, Governance, and Policies in Artificial Intelligence*; Springer: Cham, Switzerland, 2021; pp. 153–183. Available online: https://link.springer.com/chapter/10.1007/978-3-030-81907-1_10 (accessed on 23 January 2023).
3. Alter, G.C.; Vardigan, M. Addressing Global Data Sharing Challenges. *J. Empir. Res. Hum. Res. Ethics* **2015**, *10*, 317–323. [CrossRef] [PubMed]
4. Ballantyne, A. Adjusting the focus: A public health ethics approach to data research. *Bioethics* **2019**, *33*, 357–366. [CrossRef] [PubMed]
5. Dietrich, S.; van der Ham, J.; Pras, A.; Deij, R.V.R.; Shou, D.; Sperotto, A.; van Wynsberghe, A.; Zuck, L.D. Ethics in data sharing: Developing a model for best practice. In Proceedings of the IEEE Symposium on Security and Privacy, San Jose, CA, USA, 17–18 May 2014; pp. 5–9. [CrossRef]
6. Netshakhuma, N.S. Assessment of the management of student affairs records: Case of the University of Mpumalanga in South Africa. *Rec. Manag. J.* **2019**, *30*, 23–43. [CrossRef]
7. Adams, R.; Adeleke, F.; Anderson, D.; Bawa, A.; Branson, N.; Christoffels, A.; De Vries, J.; Etheredge, H.; Flack-Davison, E.; Gaffley, M.; et al. POPIA code of conduct for research. *S. Afr. J. Sci.* **2021**, *117*, 10933.
8. Thaldar, D.; Townsend, B. Exempting health research from the consent provisions of POPIA. *Potchefstroom Electron. Law J.* **2021**, *24*, 1–31. [CrossRef]
9. Thaldar, D. Research and the meaning of 'public interest' in POPIA. *S. Afr. J. Sci.* **2022**, *118*, 1–3. [CrossRef]
10. Wachter, S. Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR. *Comput. Law Secur. Rev.* **2018**, *34*, 436–449. [CrossRef]
11. Longo, D.L.; Drazen, J.M. Data sharing. *N. Engl. J. Med.* **2016**, *374*, 276–277. [CrossRef]
12. Chawinga, W.D.; Zinn, S. Global perspectives of research data sharing: A systematic literature review. *Libr. Inf. Sci. Res.* **2019**, *41*, 109–122. [CrossRef]
13. Wang, C.; Du, C. Optimization of physical education and training system based on machine learning and Internet of Things. *Neural Comput. Appl.* **2022**, *34*, 9273–9288. [CrossRef]
14. Moloi, T.; Adelowotan, M. Exploring the risks disclosed in South African technical vocational education and training colleges' annual reports. *S. Afr. J. Account. Audit. Res.* **2018**, *20*, 115–122.
15. Barczak, G.; Hopp, C.; Kaminski, J.; Piller, F.; Pruschak, G. How open is innovation research?—An empirical analysis of data sharing among innovation scholars. *Ind. Innov.* **2021**, *29*, 186–218. [CrossRef]
16. Asongu, S.A.; Odhiambo, N.M. Foreign direct investment, information technology and economic growth dynamics in Sub-Saharan Africa. *Telecommun. Policy* **2020**, *44*, 101838. [CrossRef]
17. Conrad, J.M.; Greene, K.T.; Walsh, J.I.; Whitaker, B.E. Rebel natural resource exploitation and conflict duration. *J. Confl. Resolut.* **2019**, *63*, 591–616. [CrossRef]
18. Church, S.; Rogers, E.; Rockwood, K.; Theou, O. A scoping review of the Clinical Frailty Scale. *BMC Geriatr.* **2020**, *20*, 393. [CrossRef]
19. Munn, Z.; Peters, M.D.; Stern, C.; Tufanaru, C.; McArthur, A.; Aromataris, E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med. Res. Methodol.* **2018**, *18*, 143. [CrossRef]
20. Drosatos, G.; Kaldoudi, E. Blockchain applications in the biomedical domain: A scoping review. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 229–240. [CrossRef]
21. Abebe, R.; Aruleba, K.; Birhane, A.; Kingsley, S.; Obaido, G.; Remy, S.L.; Sadagopan, S. Narratives and counternarratives on data sharing in Africa. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, Virtual Conference, Toronto, ON, Canada, 3–10 March 2021; pp. 329–341. [CrossRef]
22. Riggs, E.R.; Azzariti, D.R.; Niehaus, A.; Goehringer, S.R.; Ramos, E.M.; Rodriguez, L.L.; Knoppers, B.; Rehm, H.L.; Martin, C.L.; Clinical Genome Resource Education Working Group. Development of a consent resource for genomic data sharing in the clinical setting. *Genet. Med.* **2019**, *21*, 81–88. [CrossRef] [PubMed]
23. Bonthu, S.; Rama Sree, S.; Krishna Prasad, M.H.M. Automated Short Answer Grading Using Deep Learning: A Survey. In *Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2021; pp. 61–78. [CrossRef]
24. Hariharasudan, A.; Kot, S. A scoping review on Digital English and Education 4.0 for Industry 4.0. *Soc. Sci.* **2018**, *7*, 227. [CrossRef]
25. Ji, G.; Yu, M.; Tan, K.H. Cooperative innovation behavior based on big data. *Math. Probl. Eng.* **2020**, *2020*, 4385810. [CrossRef]
26. Barbeau, S.J.; Cetin, C. *Rapidly Expanding Mobile Apps for Crowd-Sourcing Bike Data to New Cities*; Portland State University; National Institute for Transportation and Communities: Portland, OR, USA, 2017.
27. Marivate, V.; Aghoghovwia, P.; Ismail, Y.; Mahomed-Asmail, F.; Steenhuisen, S.L. The Fourth Industrial Revolution-what does it mean to our future faculty? *S. Afr. J. Sci.* **2021**, *117*, 1–3. [CrossRef] [PubMed]
28. Park, H.; Wolfram, D. An examination of research data sharing and re-use: Implications for data citation practice. *Scientometrics* **2017**, *111*, 443–461. [CrossRef]

29. Michener, W.K. Ecological data sharing. *Ecol. Inform.* **2015**, *29*, 33–44. [CrossRef]
30. Staunton, C.; Adams, R.; Dove, E.S.; Harriman, N.; Horn, L.; Labuschaigne, M.; Mulder, N.; Olckers, A.; Pope, A.; Ramsay, M.; et al. Ethical and practical issues to consider in the governance of genomic and human research data and data sharing in South Africa: A meeting report. *AAS Open Res.* **2019**, *2*, 15. [CrossRef] [PubMed]
31. Dlamini, N.; Mazenda, A.; Masiya, T.; Nhede, N.T. Challenges to strategic planning in public institutions: A study of the Department of Telecommunications and Postal Services, South Africa. *Int. J. Public Leadersh.* **2020**, *16*, 109–124. [CrossRef]
32. Puig, F. Global Governance as Promise-Making. Negotiating and Monitoring Learning Goals in the Time of SDGs. Ph.D. Thesis, University of Barcelona, Barcelona, Spain, 2021.
33. Fischer, C.; Pardos, Z.A.; Baker, R.S.; Williams, J.J.; Smyth, P.; Yu, R.; Slater, S.; Baker, R.; Warschauer, M. Mining big data in education: Affordances and challenges. *Rev. Res. Educ.* **2020**, *44*, 130–160. [CrossRef]
34. MacPherson, Y.; Pham, K. Ethics in Health Data Science. In *Leveraging Data Science for Global Health*; Springer: Cham, Switzerland, 2020; pp. 365–372.
35. Enakrire, R.T. Data literacy for teaching and learning in higher education institutions. *Libr. Hi Tech News* **2020**, *38*, 1–7. [CrossRef]
36. Marx, M.F.; London, L.; Burnhams, N.H.; Ataguba, J. Usability of existing alcohol survey data in South Africa: A qualitative analysis. *BMJ Open* **2019**, *9*, e031560. [CrossRef]
37. Arnott, R.; Harvey, C.R.; Markowitz, H. A Backtesting Protocol in the Era of Machine Learning. *J. Financ. Data Sci.* **2019**, *1*, 64–74. [CrossRef]
38. Swales, L. The Protection of Personal Information Act and data de-identification. *S. Afr. J. Sci.* **2021**, *117*, 10808. [CrossRef] [PubMed]
39. Ross, M.W.; Iguchi, M.Y.; Panicker, S. Ethical aspects of data sharing and research participant protections. *Am. Psychol.* **2018**, *73*, 138–145. [CrossRef] [PubMed]
40. Kurze, A.; Bischof, A.; Totzauer, S.; Storz, M.; Eibl, M.; Brereton, M.; Berger, A. Guess the data: Data work to understand how people make sense of and use simple sensor data from homes. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–12.
41. Sweedler, J.V. Where Is the Data? *Anal. Chem.* **2018**, *90*, 8721. [CrossRef] [PubMed]
42. Yates, D.; Beale, T.; Marshall, S.; Parr, M. Designing data sharing agreements: A checklist. *Gates Open Res.* **2018**, *2*, 44.
43. Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; Aroyo, L.M. 'Everyone wants to do the model work, not the data work': Data Cascades in High-Stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; pp. 1–15. [CrossRef]
44. Gao, J. The Data Privacy Regulations for the Health Data in Wearable Industry in the United States. Bachelor's Thesis, Malmö University, Malmo, Sweden, 2022.
45. Elsaify, M.; Hasan, S. *Some Data on the Market for Data*; Duke Fuqua School of Business: Durham, NC, USA, 2020. [CrossRef]
46. Motloung, O.M. Governance of Digital Innovation in the Public Sector in South Africa. Ph.D. Thesis, North-West University, Potchefstroom, South Africa, 2022.
47. Sutherland, E. The Fourth Industrial Revolution—The Case of South Africa. *Politi* **2019**, *47*, 233–252. [CrossRef]
48. Chesbrough, H.W.; Appleyard, M.M. Open Innovation and Strategy. *Calif. Manag. Rev.* **2007**, *50*, 57–76. Available online: https://journals.sagepub.com/doi/pdf/10.2307/41166416 (accessed on 23 January 2023). [CrossRef]
49. McCann, C.; Talbot, A.; Westaway, A. Social capital for social change: Nine Tenths Mentoring Programme, a solution for education (in) justice in South Africa. *Int. J. Educ. Leadersh. Prep.* **2021**, *16*, 45–59.
50. Brink, C. *The Responsive University and the Crisis in South Africa*; BRILL: Leiden, The Netherlands, 2021.
51. Ruttkamp-Bloem, E. The Quest for Actionable AI Ethics. In Proceedings of the Southern African Conference for Artificial Intelligence Research, Durban, South Africa, 6–10 December 2021; Springer: Cham, Switzerland, 2020; pp. 34–50.
52. Alhakbani, H.A.; Alnassar, F.M. Open Learning Analytics: A Systematic Review of Benchmark Studies using Open University Learning Analytics Dataset. In Proceedings of the 7th International Conference on Machine Learning Technologies, Rome, Italy, 11–13 March 2022; pp. 81–86.
53. Zahle, J. Data, epistemic values, and multiple methods in case study research. *Stud. Hist. Philos. Sci. Part A* **2019**, *78*, 32–39. [CrossRef]