GREEN CLIMATE FUND | Independent Evaluation Unit

TRUSTED **EVIDENCE**.
INFORMED **POLICIES**.
HIGH **IMPACT**.

MACHINE LEARNING AND ITS POTENTIAL APPLICATIONS IN THE INDEPENDENT EVALUATION UNIT OF THE GREEN CLIMATE FUND: A SCOPING STUDY

David Huang, Byungsuk Lee, Hellen Nassuna, Martin Prowse

IEU Working Paper No. 5, 2021

Machine learning and its potential applications in the Independent Evaluation Unit of the Green Climate Fund: A scoping study

David Huang, Byungsuk Lee, Hellen Nassuna, Martin Prowse

Green Climate Fund
175, Art center-daero
Yeonsu-gu, Incheon 22004
Republic of Korea
Tel. (+82) 032-458-6450
Email: ieu@gcfund.org
https://ieu.greenclimate.fund

First Edition

This paper is a product of the Independent Evaluation Unit at the Green Climate Fund (GCF/IEU). It is part of a larger effort to provide open access to its research and work and to make a contribution to climate change discussions around the world.

While the IEU has undertaken every effort to ensure the data in this report is accurate, it is the reader's responsibility to determine if any and all information provided by the IEU is correct and verified. Neither the author(s) of this document nor anyone connected with the IEU or the GCF can be held responsible for how the information herein is used.

Citation
The suggested citation for this paper is:
Huang, David, and others (2021). Machine learning and its potential applications in the Independent Evaluation Unit of the Green Climate Fund: A scoping study. Working paper No. 5 (July). Songdo, South Korea: Independent Evaluation Unit, Green Climate Fund.

Credits
Head of the GCF Independent Evaluation Unit a.i.: Andreas Reumann
Editing: Toby Pearce
Layout and design: Giang Pham

A FREE PUBLICATION

Printed on eco-friendly paper

About the IEU

The IEU was established by the GCF Board as an independent unit, to provide objective assessments of the results of the Fund, including its funded activities, its effectiveness, and its efficiency. The IEU fulfils this mandate through four main activities:

Evaluation: Undertakes independent evaluations at different levels to inform GCF's strategic result areas and ensure its accountability.

Learning and communication: Ensures high-quality evidence and recommendations from independent evaluations are synthesized and incorporated into GCF's functioning and processes.

Advisory and capacity support: Advises the GCF Board and its stakeholders of lessons learned from evaluations and high-quality evaluative evidence, and provides guidance and capacity support to implementing entities of the GCF and their evaluation offices.

Engagement: Engages with independent evaluation offices of accredited entities and other GCF stakeholders.

About IEU Working Paper series

The IEU's Working Paper series is part of a larger effort to provide open access to the IEU's work and to contribute to global discussion on climate change. The series disseminates the findings of work in progress and encourages an exchange of ideas about climate change. Their intention is to promote discussion. The findings, interpretations and conclusions are entirely those of the authors. They do not necessarily reflect the views of the IEU, the GCF or its affiliated organizations or of the governments associated with it.

About this Working Paper

This paper conducts a scoping study of the contemporary uses of machine learning for the evaluation of climate interventions. To this end, the paper reviews the current applications of machine learning within climate impact evaluation and evidence reviews, as well as possible applications within the work of the GCF and the IEU.

About the authors

David Huang was an Evaluation Researcher at the IEU. His main duties at IEU involved performing data analyses, visualization, and supporting IEU evaluations.

Byungsuk Lee is a Research Assistant Consultant at the IEU, assisting the unit with data collection, analyses and visualization, along with conducting literature reviews and providing support for IEU evaluations.

Hellen Nassuna was an intern at the IEU, providing technical, operational and research assistance to the unit.

Martin Prowse is Evaluation Specialist at the IEU and applies his 15 years of experience in international development to support the unit on impact evaluation (LORTA), evidence reviews and behavioural science. He has published widely and is an editor of the European Journal of Development Research. Martin holds a Ph.D. from the University of Manchester.

CONTENTS

ACKNOWLEDGEMENTS

We would like to thank Joachim Vandercasteelen and Liam O'Dea for their comments on draft versions of this IEU working paper. We would also like to thank Jyotsna (Jo) Puri for encouraging us to consider the possible applications of machine learning within the works of the IEU. In addition, we would like to thank Jos Vaessen for early engagement on this topic and for facilitating one of the reviews. Any errors or inconsistencies are entirely the responsibility of the authors.

ABSTRACT

This paper explores the extent to which and how machine learning can support the evaluation function of the Independent Evaluation Unit (IEU) and, more broadly, how it can support project development at the Green Climate Fund (GCF). The paper suggests machine learning can be an effective tool for optimizing the identification and selection of beneficiaries, generating necessary data and information for proposal development that can be used within feasibility studies, and identifying the factors most relevant for project approval. The paper also suggests that machine learning can contribute to the measurement of outcomes from GCF investments, by monitoring outcomes during project implementation and filling data gaps when measuring outcomes. In addition, machine learning can improve the accuracy of specific counterfactual evaluation techniques. The paper explains the increasing role of machine learning in evidence reviews and illustrates the variety of text-mining approaches for document screening and synthesis, that can expedite the review of evidence on a specific thematic area. The paper concludes by highlighting the vital ethical dimensions of using machine learning for evaluation, and how the IEU can learn from the existing approaches to machine learning used by other multilateral agencies.

A. INTRODUCTION

The Independent Evaluation Unit (IEU) was established to provide objective assessments of the performance and results of the Green Climate Fund (GCF), thereby ensuring that the GCF is effective and is held accountable. The Updated Strategic Plan for the Green Climate Fund: 2020–2023, adopted at the twenty-seventh meeting of the Board (B.27) in November 2020, states that one upcoming priority of the GCF is to "advance as a digital organization by seeking to improve key work processes through automation" (Green Climate Fund, 2020). In this respect, this scoping study aims to contribute to two converging organizational objectives: for the IEU to become a thought leader in the evaluation of climate change interventions and for the GCF to become a digital organization. This learning product aims to inform the upcoming processes and programming within the GCF by responding to the following question: to what extent and how can machine learning support the evaluation function of the IEU and the work of the GCF?

Often used interchangeably with artificial intelligence, which is a broader term for systems that are programmed to mimic human intelligence, the term machine learning can be defined as a form of artificial intelligence that enables a system to learn from data rather than through explicit programming (Hurwitz & Kirsch, 2018). The dialogic and iterative feature of machine learning differentiates it from other multivariate statistical techniques that solely use a static data source (whether cross-sectional, time series, or panel). It is commonly categorized into (ibid.):

- Supervised learning: There is an existing set of data and understanding of how the data is classified, so the user provides the algorithm with example data and corresponding target responses that it learns from.
- Unsupervised learning: The algorithm learns by itself from example data without any corresponding data responses with the aim of understanding the meaning based on the patterns and clusters it finds.
- Reinforcement learning: The algorithm is not trained with any external datasets but learns by reacting to its own environment through trial and error.

In the past, the use of machine learning has been limited due to technological constraints. Recent advances in processing power and data storage capacity have resulted in growing applications across many spheres (Huntingford, et al., 2019). The increase in the quantity and quality of data has been conducive to the growth of machine learning as the algorithms are more effective when trained upon or applied to larger datasets (ibid.).[1] Such growth has been observed in both developed and developing countries, where it is utilized in a range of applications from data collection and processing (e.g. supervised learning of satellite data) to data analyses (e.g. making model-based predictions) and in a range of disparate areas, from determining credit scores to measuring job performance and predicting the likelihood of crime.

The growth in machine learning applications has also been observed in the area of climate change, which is particularly data-intensive (ibid.). The emergence of this nascent field has been buoyed through recent advances in technology[2] and data storage capacity.[3] Machine learning sets itself apart

[1] Training refers to the process by which the algorithm learns from the data and subsequently builds a model to provide the necessary output, usually in the form of regression or classification.
[2] For example, neural network algorithms have been around for decades, yet it is only the recent improvements in computational architecture and power that have led to its growth in practical applications. Similarly, climate science has also benefited from recent technological advancements which have resulted in high resolution weather forecasting and climate projection models (Huntingford, et al., 2019).
[3] The rise of big data (large amounts of digitally generated data from social media, online searches, financial transactions, etc.) has allowed for the efficient collection of data that would otherwise be very costly or time consuming to collect for impact evaluations (Rathinam, et al., 2020).

from traditional statistical methods by its ability to model complex data across scales, as well as to model the fundamental uncertainty typical of climate change. It is likely to be more accurate than traditional statistical methods and computationally less costly (Rolnick, et al., 2019). Moreover, recent research has demonstrated the use of machine learning methods for "targeting, monitoring, and evaluating efforts to mitigate climate change" (Leo, et al., 2020). Machine learning can be useful for evaluations of climate change interventions, with the availability of extremely large and granular satellite data and the replicability of the approach across regions or scales. In addition, the immobility brought about by the ongoing COVID-19 pandemic presents an opportune time to explore new ways of analysing and making sense of existing data for climate change interventions and evaluations.

This paper presents the potential applications of machine learning in four areas of the work of the GCF, including project development, results measurement, evaluations and evidence reviews. Each section describes potential applications and their existing examples taken from both peer-reviewed and grey literature. It should be noted that many of the examples discussed are at the proof-of-concept stage, but they still show how machine learning in principle could be applicable to the work of the GCF and the IEU. The last section concludes the paper and discusses the ethical dimensions of using machine learning.

B. USE OF MACHINE LEARNING IN PROJECT DEVELOPMENT

We now turn to the first section, which discusses the potential use of machine learning in project development before implementation, which takes up a large portion of the GCF's work. The GCF puts great emphasis on the preparation and development of the projects to ensure, ex ante, the climate change relevance and success of the projects.

Machine learning could be an effective tool in optimizing the identification and selection of beneficiaries to be supported by the intervention of a potential project (Mckenzie, 2018). When information on the characteristics of potential beneficiaries or regions is missing, which is a determining factor for inclusion in the project intervention, machine learning and predictive modelling in particular, could generate the missing information based on other existing relevant information. This also applies to data in the future which can be extrapolated from current information sources.

In such predictive modelling, it is important to identify the most relevant variables pertaining to the potential beneficiaries as predictors for the variable that is missing or not available. When information on the determining factor is not readily available, the related variables can act as either predictors or proxies for the determining factor. An example of this is found in a case study from Malawi, where a linear regression and a "random forest" algorithm[4] were used to identify that, among existing factors, living in the flood plain and distance to drinking water were consistently good predictors for food insecurity (Knippenberg, Jensen, & Constas, 2018). In other words, food insecurity can be predicted using the two variables and the predictive models aiding the selection of beneficiaries.

Next, in addition to optimizing the selection of beneficiaries, machine learning could help provide the information and data required for the development of projects that are otherwise too costly for

[4] Random forest is a machine-learning algorithm that fits multiple decision trees to input data using a random subset of the input variables for each tree constructed; the mode (or average) of these trees is used to create an "ensemble" tree that is used for prediction. In this study, multiple trees were created from a randomly selected subset of variables, and each tree was a binary tree where each pair of branches was defined by finding a variable (among the selected variables) and a value (of the variable) that minimize the mean-squared error of the dependent variable (which in this case is the Coping Strategy Index used as a measure of food insecurity). The branches are then defined such that one consists of observations whose value of the chosen variable is greater or less than the chosen value.

developing countries to obtain. It holds in many cases that "countries which are disproportionately affected by climate change are often the world's data-poor as well" (Leo, et al., 2020). The recent IEU evaluation of the GCF portfolio in small island developing states highlighted the high human resource and financial costs of preparing a successful funding proposal. For many organizations in small island developing states, the submission of mandatory feasibility studies within project proposals is challenging due to limited data availability, human resources, and the time and financial costs of completing this requirement. Machine learning could play a role in situations like this. For example, a recent study in Bangladesh showed how machine learning helped facilitate the identification of the climate-related agricultural vulnerabilities of individual households for vulnerability assessments (Jakariya, et al., 2020). The study used machine learning algorithms to identify the most relevant factors to measure vulnerability and develop a predictive model. Then, using a mobile application with an easy-to-use user interface, they collected the inputs from smallholder farmers, which were uploaded to a server where the regression model calculated the vulnerability score for each individual household.[5]

Another way in which the GCF could benefit from using machine learning for feasibility studies and climate change vulnerability assessments is by studying textual data from local and social media. Methods of text analytics could be applied to online discussions on climate risks and vulnerabilities, to identify which factors are consistently negatively mentioned, For example, a study from Italy used tweets (textual data) gleaned from Twitter to train a supervised sentiment analysis classifier to analyse active users' opinions on certain public administration decisions in the country (Corallo, et al., 2015). Similarly, the GCF could take advantage of the existing information in the local and social media of a country to contextualize the extent of the country's climate vulnerabilities, the needs of potential beneficiaries, or even public perception of the GCF and relevant accredited and executing entities, without physically visiting or interviewing the country's stakeholders. Such an undertaking would imply the existence of active local and social media that frequently discuss the country's climate issues.

Lastly, predictive modelling in machine learning could be used to identify predictors for the success of GCF funding proposals, based on the data of previously approved and implemented funding proposals. Potential features or predictors of such models include any quantifiable characteristic of the project, be it geographical or financial, including its size, amount of GCF funding and co-financing, co-financers, its accredited entity (AE) and duration. Once the most relevant predictors are identified and the predictive models developed, they could be used to predict the likelihood that a potential project is approved by the Board. This could be applicable to the GCF project pipeline, which would allow GCF staff to advise project developers to course-correct, and to adjust their proposals. A range of options for such predictive models exist, such as those which use maximum likelihood estimates based on a probit or logit model. However, for any of the models to be effective in use, a considerable amount of training data would be necessary (which would detail the project proposals that have been developed, and show which of these have been submitted to and approved by the Board). Due to the relative youth of the GCF as an institution, it may be the case that an insufficient number of project proposals are available. If this is the case, one alternative may be to use the project information from a comparable climate institution where project proposals may have the same information (predicators) as for GCF projects.

[5] Namely, linear regression and Bayesian ridge regression. The study also tested random forest, extreme gradient boosting, and extremely randomized tree regressions, but eventually selected the aforementioned two due to their simplicity (and lack of necessity for hyperparameter tuning).

C. USE OF MACHINE LEARNING IN IMPACT EVALUATION AND THE MEASUREMENT OF OUTCOMES

We now discuss the potential use of machine learning in impact evaluation, including in the measurement of project outcomes. The IEU's Learning Oriented Real-Time Impact Assessment (LORTA) programme embeds impact evaluation techniques within projects' monitoring and evaluation activities, to measure the impact of GCF projects on the ground. A growing body of research is showing that machine learning could "revolutionize impact evaluations" and is "especially powerful [in] evaluating efforts to mitigate climate change" (Leo, et al., 2020, p. 24). Impact evaluations measure the changes created by interventions. They take the form of either experimental or quasi-experimental designs. Experimental designs involve evaluating the impact of a treatment or intervention through randomizing the selection of beneficiary and comparison groups. This controls for confounding factors such as selection bias. Impact is ascertained by comparing the outcomes of both these groups. Given that randomization should ensure that both populations share the same observable and unobservable characteristics, the difference between the outcomes of the two populations can be attributed to the impact of the intervention. However, in many instances, it is not possible to implement such an experimental design for ethical, practical, or technical reasons. In the absence of a control population, a counterfactual comparison group may be generated through the use of quasi-experimental designs. Such designs use statistical means to replicate the conditions generated through randomization.[6]

Given that most machine learning techniques are predictive in nature, they could be perceived as being unsuitable for causal inference and consequently are underutilized in quasi-experimental designs, with econometric models traditionally being the preferred choice for evaluators (Dance & Hawksworth, 2017). However, recent work on the use of machine learning to optimize one quasi-experimental technique, namely propensity score matching, illustrates the application of machine learning in this area. Propensity score matching is a counterfactual technique that generates the probability of participation in the project intervention by generating a propensity score for each observation (individual or group) in the total population based on selected predictor variables. The generation of the propensity score is typically done with a logit or probit model, but it could also benefit from using more complex machine learning algorithms such as random forest and neural networks (Cannas & Arpino, 2019). Treated observations are then matched (using a variety of matching approaches) with non-treated counterparts who share very similar propensity scores.[7]

Another counterfactual matching technique that is simpler and commonly used is a form of direct matching, called "direct nearest neighbour matching". Instead of calculating another metric such as a propensity score from observations, direct nearest neighbour matching simply matches the observations that are closest to each other (using n-dimensional Euclidean distance). This approach does have some downsides. For example, its precision declines with the dimensionality of the covariates (in other words, the number of predictor variables). Hence, its accuracy could be improved by employing dimensionality reduction (Li, Vlassis, Kawale, & Fu, 2016). Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality (van der Maaten, Postma, & van den Herik, 2009). It is important and widely used

[6] For example, difference-in-difference designs use panel data to control for confounding factors, while propensity score matching creates an artificial comparison group using maximum likelihood estimations. These are then twinned with beneficiary units. Two further approaches are regression discontinuity designs, which use the threshold for eligibility to compare units just either side of the cut-off point, and instrumental variable regression, which uses a variable that 'cleans' the treatment variable of any bias.

[7] Observations are matched between those who are within the range of common support – that is, the range of propensity scores which feature in both treatment and control groups. More detailed explanation and application can be found in (Berhane & Prowse, 2013).

because it mitigates the requirement of having enormously big training data for the high dimensionality of features (Jimenez & Landgrebe, 1999). A range of dimensionality reduction techniques exist, such as principal component analysis and locality preserving projection, and they could be employed through machine learning.[8]

Machine learning could also be applied to measuring the outcomes or impact of an intervention, which is an integral part of impact evaluation. When direct outcome measurements are difficult or insufficient, predictive modelling could play a role in filling the gap or replacing direct measurements. For example, in order to measure the impact of infrastructure investments that support urbanization, predictive models could be used to analyse the urbanization rates from satellite imagery (after being trained together with another separate dataset that indicates urban versus non-urban areas) (Goldblatt, You, Hanson, & Khandelwal, 2016). A further example concerns measuring changes in socioeconomic indicators such as wealth. A predictive model based on deep learning could be developed using nightlight satellite imagery and a manually constructed geographical wealth index to generate geospatial distribution of wealth in the area of interest (Yeh, et al., 2020). In both examples, predictive models could support the measurement of outcomes such as urbanization rates or changes in the level of wealth from satellite data, when ground measurements prove difficult or insufficient.

Moreover, machine learning could be used to enable the measurement of outcomes that would otherwise be difficult to ascertain (Mckenzie, 2018). Textual analytics using machine learning could allow for automatic detection of patterns and trends within large volumes of textual data that are difficult to systematically analyse using traditional means. For example, sentiment analysis could be applied to large amounts of unstructured textual data – such as transcripts from radio show discussions – to understand the effects of recently implemented policies (UN Global Pulse, 2017).

For the effective measurement of outcomes, it is vital to track and monitor the project during implementation. Here, image processing techniques using machine learning coupled with remote sensing technologies could again play a role. For example, a mitigation project may require carbon mapping for forestry monitoring and tracking of carbon emissions. Machine learning could be employed in such monitoring (Mascaro, 2014). Measuring the amount of carbon stored in a patch of forest is, in theory, relatively straightforward, but traditional means of measurement have proved inadequate for estimating the spatial distribution of carbon stocks at a larger scale. Hence, remote sensing technologies such as LiDAR have been used to capture forest geospatial data and subsequently estimate the spatial variation of the carbon stocks, sometimes more accurately than traditional methods. However, their application is also limited by cost and logistical considerations, and as such they are often only used on a limited geographical scale. Here, Mascaro (2014) applied random forest algorithms[9] to a variety of remote sensing data (such as Landsat) with greater geographical coverage to spatially upscale (i.e. increase the geographic coverage of) the smaller-scaled LiDAR-based carbon maps. It found that while there were some drawbacks, random forest offers a viable way to upscale carbon mapping over a larger geographic area.

Given that accurate carbon mapping is essential for the monitoring and evaluation of forestry-related mitigation projects, such application may be highly pertinent to the GCF. Currently, GCF projects are monitored through the submission of annual performance reports, which are provided by the AEs of the projects. Much of the project monitoring data found within the annual performance reports are self-reported by the entities. Recent evaluative work by the IEU found that the use of

[8] An example of a traditional approach to using principal components analysis in relation to climate adaptation can be seen in Mahmud & Prowse (2012).
[9] In this paper's case, the input variables used were geographic properties (e.g. coordinates and elevation) and land use estimations (e.g. percentage cover of soil or vegetation) for each pixel in different satellite images (after resampling to the carbon map's resolution).

such data makes it difficult to draw substantive conclusions on the realized results of the GCF's projects, due to lack of guidance on the methodology as well as the GCF's due diligence on the reporting standards of entities. Predictive modelling as described above could prove to be a more accurate and standardized way to monitor the carbon stocks of GCF project sites and verify the resulting emission reductions.

D. USE OF MACHINE LEARNING IN SYSTEMATIC EVIDENCE AND LITERATURE REVIEWS

We now turn to the potential use of machine learning in evidence and literature reviews, which are a critical component of not only the IEU's ongoing evaluation work, but also the knowledge management of the GCF as a thought leader in climate change actions.

An evidence review is a comprehensive, systematic and rigorous collation, analysis and presentation of evidence. It may be either qualitative or quantitative in nature and is often thematic with the aim of assessing the evidence base on a topic. It provides a concise overview of available findings dispersed across numerous disciplines and sources, and is based on a structured literature search guided by a protocol. It offers an appraisal of the quality of evidence based on clear criteria and provides an analytical synthesis of the evidence base. The evidence itself can take a myriad of forms and may include both peer-reviewed journal articles and grey literature (e.g. reports published by not-for-profit organizations and international development organizations). Products from evidence reviews include evidence gap maps or systematic reviews.

An evidence gap map graphically depicts a collection of evidence across a range of interventions and outcome categories in a thematic area. Its overarching framework is guided by a theory of change, predefined inclusion and exclusion criteria, and the generation of an intervention/outcome framework. Both impact evaluations and systematic reviews (which we come to below) may be mapped onto an evidence gap map in addition to individual articles and papers. Evidence gap maps highlight the existing evidence base, or lack thereof, on the effectiveness of interventions towards different outcomes. Importantly, they highlight gaps in evidence, which may take one of three forms. First, there may be little evidence of any type on the contribution of a particular intervention-outcome combination. Second, there may be evidence but few causal designs (such as experimental or quasi-experimental approaches) which are able to attribute change within outcome areas to specific interventions. These gaps provide clear areas within which impact evaluations are needed. Third, there may be a wealth of impact evaluations for a particular intervention/outcome cell, but few systematic reviews that aggregate salient findings from the available evidence base. In this way, evidence gap maps can be used as a precursor for systematic reviews (which we turn to shortly). Ultimately, evidence gap maps can meaningfully inform policy decisions and strategically utilize finite research funding.

Systematic reviews attempt to synthesize available evidence on a given topic. Their scope is necessarily narrow, and they almost exclusively focus on causal studies. They are always guided by a strict protocol. As with evidence gap maps, systematic reviews are subject to a set of inclusion and exclusion criteria which determine the scope of studies that are included. A search strategy guides study selection, which is subsequently screened for quality. Articles are assessed and categorized independently by reviewers and subsequently analysed. Where possible, quantitative evidence is subject to meta-analysis, while qualitative evidence is subject to theory-based analysis. At completion, the report is subject to a peer-review process, which ensures that the methodology is valid and relevant conclusions are sound. Systematic reviews are often assessed using the Specialist Unit for Review Evidence (SURE) model, and are subsequently coded in line with the level of

confidence associated with their conclusions and the effects of particular interventions (low, medium, or high).

The ever-expanding pool of research and data in many academic fields, both natural and social, is surpassing the capacities of researchers and scientists to thoroughly examine, manage and utilize the evidence (Hey & Trefethen, 2003). This can put a strain on producing a review of the literature in a field, and it becomes harder and harder to check every relevant article or publication. The fields of interest of the IEU, such as climate evaluation, low-emission pathways and climate-resilient development, are no exception to this phenomenon. Put simply, there are often just too many relevant studies for evaluators to keep up to date with.

The conventional practice when completing a literature review is to use keyword searches to gather a volume of potentially relevant studies, and then manually screen them. This is time-consuming as it may include tens of thousands of titles and abstracts. For a more efficient literature review to take place, it is now becoming more common to consider employing a range of text mining tools based on natural language processing (NLP) that can automate the process of dealing with large amounts of textual data, thereby saving time and improving the effectiveness and efficiency of these types of reviews. One of the clearest examples is systematic reviews. These consist of the following three steps: searching, screening, and then synthesizing (Ananiadou et al., 2009). Searching involves using a query in an electronic database or online to locate as many potentially relevant studies and references as possible. Screening narrows down the resources collected from searching to a smaller list that is relevant to a specific topic. Synthesizing combines the key elements and results from the narrowed-down list to produce a summarized review. The application of text mining to systematic literature reviews can support all the three steps described above in different ways (ibid.).

First, searching can be improved using query expansion techniques (ibid.). Query expansion indicates the process of expanding a given query in a search to include more relevant terms that can lead to better matching search results. Query expansion involves automated techniques that aim to find synonyms, acronyms, semantically related words, various orthographic and morphological forms of words, and spelling corrections. Developing a query expansion model to map technical terms to their variants and other relevant terms, could enhance the search results for the initial search procedures within the IEU's systematic reviews.[10]

Second, screening can be improved by employing document clustering (ibid.). Document clustering is the application of an unsupervised learning algorithm for grouping documents' textual data. It creates clusters, each of which represents a topic independent of one another. Documents that are considered to have similar contents will be grouped together. Once the topics (clusters) are identified through visualization, they serve as indicators for relevance and priority in document screening, thereby reducing the overall workload. Clusters act as categories for document classification which can be employed by a supervised learning algorithm for classification, with the training dataset as the documents and their identified clusters. The classification model can be used to classify new documents, expediting the review of new documents. Document classification categories can also be manually created according to the reviewer's interest. While manually creating a training dataset may take a long time and be a project in its own right, it may prove effective in building a classification model that caters to the specific needs of the reviewer.

Document clustering can also provide a way for screening prioritization. Screening prioritization provides an ordered list of items, with the most likely relevant items at the top of the list (O'Mara-Eves et al., 2015). Given a document that is most relevant to the reviewer's interest, document clustering can provide an ordered list of documents based on their resemblance (or computed

[10] This would require the IEU to maintain its own database of literature and develop a search engine of its own for the database.

distance) to the initial document. Several benefits exist with this method when compared to random document screening. One is that the reviewer encounters more relevant examples earlier, which speeds up determining which documents to dive into for more in-depth screening and information retrieval. This can also help with document searching, because identifying relevant bibliographies early on leads to faster searching for more relevant new external documents. Another benefit is that, if a threshold for the number of items to be screened exists, screening prioritization is a good criterion by which to enforce the threshold, thereby saving workload. This also prevents a common problem of over-inclusiveness that researchers face, trying to review as many documents as possible through the fear of missing critically relevant studies.

Third, synthesizing can be improved by employing automatic summarization (Ananiadou et al., 2009). Single-document text summarization has two general approaches: extraction-based summarization and abstract-based summarization (Hovy & Lin, 1999). Extract-based summarization identifies and extracts key, important sentences and phrases from the original text. Abstract-based summarization involves using semantic representation of words in the original text to create a new, paraphrased summary. Abstract-based summarization requires first developing a library of semantic linkages for technical terms, which may require specific, expert knowledge in the reviewer's topic. For the purpose of a systematic review, multi-document summarization can be employed using sentence-ordering algorithms combined with text summarization techniques (Barzilay, Elhadad, & McKeown, 2001). In multi-document summarization, the key is to produce a coherent arrangement of summary sentences and phrases from the individual documents. The benefit of multi-document summarization is that, ideally, the key information is already present in the resulting summary, reducing the need to manually examine the original individual documents.

Developing a text-mining model as described above would entail the following steps, which are common to many NLP applications (Ng, n.d.):

1) Tokenization: Parsing textual data from documents into smaller tokens such as words or short phrases. Example methods include the bag-of-words model and n-gram model.
2) Stemming/lemmatization: Processing tokens such that they are reduced to either their stems (the least common spelling denominator) or lemmas (the base dictionary form for a group of word inflections).
3) Removing stop words (such as "the", "a", "this", etc.) and punctuation.
4) Generating features: A common way to generate features of text data is computing the term frequency of the tokens. Sometimes the frequencies are weighted using techniques such as term frequency-inverse document frequency (tf-idf in short) to give greater emphasis on technical or topic-specific terms and less on general terms that frequently appear throughout different documents.

One common issue in developing a machine learning model for information retrieval or classification is the trade-off between precision and recall (Ng, n.d.). Considering a test dataset with items marked as either positive or negative and a model predicting a positive or a negative for each input item, precision is the percentage of true positives in the predicted positives, while recall is the percentage of true positives in the actual positives in the test dataset. Having a higher threshold in the classifier tends to result in higher precision and lower recall, while having a lower threshold in the classifier tends to result in lower precision and higher recall. This trade-off would be determined depending on the purpose of the model. In document classification, if the purpose is to precisely narrow down the collected literature to a very specific topic, having a higher precision may be favoured. However, if the purpose is to secure as many relevant documents to a topic as possible, however vague, then having a higher recall may be preferred. If both precision and recall are

concerned, then the harmonic mean of precision and recall, known as the "F1" score, could also be used to gauge the effectiveness of the model.

The IEU has already employed existing software in one of its evidence reviews, namely EPPI-Reviewer 4 (Thomas, et al., 2020), to reduce the workload in identifying relevant documents for the evidence review's inclusion criteria. The EPPI-Reviewer 4 provides a "priority screening" function, which employs a machine learning algorithm to prioritize a given number of studies according to relevance based on titles and abstracts. It uses "active learning", an iterative process where the machine's accuracy for prioritization is improved at each iteration with manual screening. At the first iteration, a number of studies are manually screened for inclusion versus exclusion and then fed into the machine as the training dataset, which then prioritizes the remaining studies by assigning a score to each one using a support vector machine algorithm. At each subsequent iteration, reviewers manually screen the documents for inclusion versus exclusion (~25 studies), which are then added to the training dataset, and the machine then prioritizes the remaining documents. This process repeats until a desired number of studies are included or the rate of inclusion of studies reaches a de facto plateau. Hence, the machine's prioritization helps the reviewers identify the relevant documents faster than manually screening all of them. While this may improve the speed and accuracy of conducting evidence reviews, a human element should be retained for quality assurance, available to "come in when discretion or expert judgement is needed" (White, 2019).

Much of IEU's work in evaluating the GCF's performance and processes requires references to the relevant Fund documentation literature and comparable documentation within similar climate funds, such as the Global Environment Facility (GEF), the Adaptation Fund (AF), and the Climate Investment Funds (CIF). In this respect, employing a text-mining tool such as document clustering, prioritization, and/or classification for better document screening would be beneficial for the IEU when benchmarking against other climate funds. Moreover, given the IEU's commitment to producing data-driven evaluations, it is only natural that data extraction and updates are common processes within the IEU. During each evaluation, raw data is extracted from unstructured sources such as the GCF's funding proposals. This data is subsequently coded, stored, and updated in a dataset. The use of machine learning algorithms can help alleviate the manual, time-extensive aspects of the extraction process by automating it.

One example where the GCF employed text mining on the funding proposals is found within the work of the Independent Integrity Unit (IIU). The IIU employed machine learning in the ranking of projects within the GCF's portfolio as part of Integrity Due Diligence, one of many reasonable steps which might be taken by an anti-fraud unit, especially in support of a decision to select projects for a Proactive Integrity Review (PIR)[11]. The IIU created a ranking system based on sentence-level red-flag detection operating on two layers: A "relevance" layer in which a machine learning model was trained to distinguish control-related sentences from sentences unrelated to controls, followed by a "performance" layer, in which a separate machine learning model was trained to distinguish sentences indicating "weak" control performance from sentences indicating "strong/normal" control performance. This system is now used to assess GCF funded activities. Following a human review of these outputs, a set of red flags is produced.

E. CONCLUSION AND FURTHER CONSIDERATIONS

In this scoping study, we have described the potential applications of machine learning for the GCF and the IEU. Considering that climate change interventions and impact evaluations are data-

[11] A PIR, which is not an investigation, contributes toward enhanced disclosure on the control performance of a counterparty with respect to a funded activity, offering the Fund a risk-reduction (or loss-control) option for fraud risk and reputation risk. Personal communication: Liam O'Dea.

intensive, machine learning is an innovative tool to supplement or replace parts of the processes in order to achieve better automation and accuracy.

First, in relation to GCF project development, machine learning was suggested as an effective tool for optimizing the identification and selection of beneficiaries, generating the data and information necessary for proposal development (such as feasibility studies), and identifying factors most relevant to project success and the prediction of pipeline project success. Next, in impact evaluation and measurement of outcomes that are of great interest to the IEU, machine learning was suggested as a way of improving counterfactual approaches, for filling data gaps in outcome measurements, enabling outcome measurements themselves, and as a method for monitoring outcomes during project implementation with a focus on carbon stocks. Lastly, in systematic evidence and literature review, a variety of text mining methods in document screening and synthesis could expedite review processes.

The examples and insights illustrated so far have the potential to contribute to the future of the GCF and the IEU. For the IEU especially, machine learning may be the beginning of a new and exciting journey, pushing the boundaries of how climate change impact evaluations are conducted. Moving forward, this may require the IEU to strengthen collaboration with data scientists and enhance partnerships with experienced data hubs, bringing in more experts to guide on best practices and formulate case studies far beyond this paper's discussion.

Before proceeding to invest in implementing the machine learning applications, however, consideration must be given to the ethical issues that surround the use of machine learning. Given that machine learning models are often complex and opaque, they can be seen as "black box" models. In the most extreme cases, even the developers of the models may be at a loss as to how to explain the way they function (which is precisely what occurred with the securitization of mortgage bonds which precipitated the financial crisis of 2008/2009). This can lead to a lack of transparency and subsequently a lack of accountability for decisions that are made based on such models. For example, if a machine learning algorithm was used to determine the targeting of a population or location to receive a treatment or intervention, groups that are excluded may feel aggrieved at not only their omission but also not knowing the grounds for their omission (Andini et al., 2018). Concerns over transparency and accountability are particularly applicable to the GCF's works because it is a publicly funded organization. Use of overly complex models to inform its decision making may be seen as an attempt to obfuscate its internal processes and shirk away from its responsibility to operate transparently.

The machine learning community is aware of the unfairness and opaqueness that could arise from using these algorithms and has made efforts to alleviate them, one of which is developing explainer libraries. A popular example is the Local Interpretable Model-agnostic Explanations (LIME) library (Lundberg & Lee, 2017). This library aims to enable an understanding of the rationale behind the model's predictions by providing insight into a machine learning model's internal workings. This may involve identifying features that had the most significant impact in predicting the result of the model, as well as sensitive features that bring about the unfairness of the model.

Another way to ensure transparency is to use models that are compiled in open-source software – to allow numerous eyes to verify the models and flag problems – and for developers to fix them early on, eventually building confidence in the audience. Similarly, any application of machine learning within the GCF could be based on comprehensive consultations with Board members and stakeholders, including civil society organizations and members of the broader climate community, to ensure that the purpose and inner workings of the models are well understood even before development. A further and tried-and-tested route to ensuring the robustness of machine learning models with social and environmental data is triangulating their findings with a range of further sources of data, such as focus groups, key informant interviews, and further qualitative methods.

Ground truthing the results from such models with the knowledge and expertise of participants, stakeholders and wider actors, is a vital aspect of validating the results from such models.

In addition to ethical issues, capacity concerns also need to be addressed before embarking on the use of machine learning. In many instances, evaluators are not necessarily machine learning experts and machine learning experts are not necessarily evaluators. However, expertise of both is essential in coming up with practical applications of machine learning for climate change initiatives and workflows. Therefore, institutional capacity building in machine learning (e.g. training current staff based on resources - see USAID, 2021), or collaboration with external practitioners (e.g. partnership with research labs, universities, firms) is an important next step.

The IEU can also learn from the work of comparator organizations that are already employing machine learning to complement their processes. In 2016, the Independent Evaluation Office of the GEF, in partnership with geospatial specialists from AidData at the College of William & Mary, conducted a study to evaluate the impact of the GEF's land degradation projects (Independent Evaluation Office, 2016). Central to this study was the use of a machine learning algorithm trained upon copious amounts of satellite imagery to assess land degradation at a very fine geospatial resolution. Land degradation of an area was calculated by applying the United Nations Convention to Combat Desertification monitoring framework, which is based on indicators such as forest cover change and vegetation productivity. The GEF's project sites were geocoded and then assessed using this algorithm, where the algorithm provided a counterfactual through predicting the land degradation in the absence of the GEF's interventions. The study found overall that GEF projects have a positive impact on "vegetative productivity and forest cover", thereby helping to mitigate land degradation. Furthermore, while the average GEF project costs USD 4.2 million, they sequester around USD 7.5 million worth of carbon. This study was the first of its kind to adopt a multi-disciplinary approach, combining machine learning, geospatial analysis, and econometric modelling to ensure a robust data-driven impact evaluation of environmental protection projects (Patterson & Custer, 2017).

A further example is the International Fund for Agricultural Development (IFAD), where a multidisciplinary team used machine learning to extract insights by applying text mining and modelling to documents accumulated for its entire investment portfolio (IFAD, 2021). With this approach, they were able to extract words associated with the Sustainable Development Goals (SDGs), and use them to visualize a time trend for the presence of the SDGs in IFAD's documents against project approval dates. They were also able to text mine and detect the main interventions mentioned for their top five sectors. These visualizations were especially helpful in relaying a comprehensive picture of IFAD's portfolio to enable targeted interventions and strategic objectives.

This paper aims to contribute to two converging organizational GCF objectives, the first of which is for the IEU to increase its use of machine learning to support its desire to become a thought leader in the evaluation of climate change interventions. The paper also aims to contribute to the goal of the GCF becoming a digital organization. The paper has found that there is ample scope for machine learning to support the IEU and the GCF, and that through appropriate capacity building and learning, the IEU and the GCF could implement these measures in the current GCF–1 programming period and support the replenishment of the Fund for the GCF–2 programming period.

REFERENCES

Ananiadou, S., Rea, B., Okazaki, N., Procter, R., & Thomas, J. (2009). Supporting Systematic Reviews Using Text Mining. *Social Science Computer Review* 27(4), 509-523.

Andini, M., Ciani, E., de Blasio, G., & D'Iganzio, A. (2018). Effective policy targeting with machine learning. Retrieved from voxeu.org: https://voxeu.org/article/effective-policy-targeting-machine-learning

Barzilay, R., Elhadad, N., & McKeown, K. R. (2001). Sentence Ordering in Multidocument Summarization. *The first international conference on human language technology research (HLT '01).* Association for Computational Linguistics. doi:10.3115/1072133.1072217

Weldegebriel, Z. B., & Prowse, M. (2013). Climate-change adaptation in Ethiopia: to what extent does social protection influence livelihood diversification?. Development Policy Review, 31, 35-56.

Cannas, M., & Arpino, B. (2019). A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biometrical Journal*, 1049-1072.

Corallo, A., Fortunato, L., Matera, M., Alessi, M., Camillò, A., Chetta, V.,... Storelli, D. (2015). Sentiment analysis for government: An optimized approach. *International Workshop on Machine Learning and Data Mining in Pattern Recognition* (pp. 98-112). Springer.

Dance, H., & Hawksworth, J. (2017, July 31). *What can machine learning add to economics?* Retrieved from PWC - UK blogs: https://pwc.blogs.com/economics_in_business/2017/07/what-can-machine-learning-add-to-economics-.html

Goldblatt, R., You, W., Hanson, G., & Khandelwal, A. K. (2016). Detecting the Boundaries of Urban Areas in India: A Dataset for Pixel-Based Image Classification in Google Earth Engine. *Remote Sens., 8*(634). doi:10.3390/rs8080634

Green Climate Fund. (2020). *Updated Strategic Plan.* Retrived from greenclimate.fund: https://www.greenclimate.fund/document/updated-strategic-plan-green-climate-fund-2020-2023

Hey, T., & Trefethen, A. (2003). The data deluge: An e-science perspective. Grid computing: Making the global infrastructure a reality, 809-824.

Hovy, E., & Lin, C.-Y. (1999). Automated Text Summarization in SUMMARIST. In I. Mani, & M. Maybury, *Advances in Automatic Text Summarization.* MIT Press.

Huntingford, C., Jeffers, E., Bonsall, M., Christensen, H., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters, Volume 14, Number 12*.

Hurwitz, J., & Kirsch, D. (2018). *Machine Learning For Dummies.* Hoboken, NJ, USA: John Wiley & Sons, Inc.

Independent Evaluation Office, Global Environment Facility. (2016). *Value for Money Analysis for the Land Degradation Projects of the GEF*. Washington, DC: Global Environment Facility.

International Fund for Agricultural Development. (2021). *Innovation Challenge: Leveraging Artificial Intelligence and Big Data for IFAD 2.0*. Retrieved from International Fund for Agricultural Development (IFAD): https://www.ifad.org/innovation-challenge/page6.html#content5-4g

Jakariya, M., Alam, M. S., Rahman, M. A., Ahmed, S., Elahi, M. L., Khan, A. M.,... Sayem, S. M. (2020). Assessing climate-induced agricultural vulnerable coastal communities of Bangladesh using machine learning techniques. *Science of the Total Environment*, 140-255.

Jimenez, L., & Landgrebe, D. (1999). Supervised Classification in High Dimensional Space: Geometrical, Statistical and Asymptotical Properties of Multivariate Data. *IEEE Transactions on Geoscience and Remote Sensing, 37*(6).

Knippenberg, E., Jensen, N., & Constas, M. (2018). *Resilience, Shocks, and the Dynamics of Food Insecurity Evidence from Malawi.* Retrieved March 22, 2021, from https://www.erwinknippenberg.com/research/

Leo, B., Pattni, S., Winn, C., Quinn, L., Paton, C., & Persaud, M. (2020). Using machine learning for climate impact evaluation. *eVALUation Matters Second Quarter 2020*.

Li, S., Vlassis, N., Kawale, J., & Fu, Y. (2016). Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 3768-3774). IJCAI 2016.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017).* Long Beach, CA, USA.

Mascaro, J. B. (2014). A tale of two "forests": Random Forest machine learning aids tropical forest carbon mapping. *PloS one*, e85993.

Mckenzie, D. (2018). *How can machine learning and artificial intelligence be used in development interventions and impact evaluations?* Retrieved from worldbank.org: https://blogs.worldbank.org/impactevaluations/how-can-machine-learning-and-artificial-intelligence-be-used-development-interventions-and-impact.

Mahmud, T., & Prowse, M. (2012). Corruption in cyclone preparedness and relief efforts in coastal Bangladesh: Lessons for climate adaptation?. Global Environmental Change, 22(4), 933-943.

Ng, A. (n.d.). *Machine Learning by Stanford University*. Retrieved from Coursera: https://www.coursera.org/learn/machine-learning

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews. *Systematic Reviews*. 4 (1), 1-22.

Patterson, S., & Custer, J. (2017, January 11). *Using machine learning to combat environmental degradation on a global scale*. Retrieved from Aiddata.org: https://www.aiddata.org/blog/using-machine-learning-to-combat-environmental-degradation-on-a-global-scale

Rathinam, F., Khatua, S., Siddiqui, K., Malik, M., Duggal, P., Watson, S., & Vollenweider, X. (2020). *Using big data for evaluating development outcomes: a systematic map.* Centre of Excellence for Development Impact and Learning. Retrieved from cedilprogramme.org: https://cedilprogramme.org/funded-projects/programme-of-work-1/using-big-data-for-evaluating-development-outcomes-a-systematic-map/

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K.,... others. (2019). Tackling Climate Change with Machine Learning. *arXiv preprint arXiv:1906.05433*.

Thomas, J., Graziosi, S., Brunton, J., Ghouze, Z., O'Driscoll, P., & Bond, M. (2020). *EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis*. (EPPI-Centre Software. London: UCL Social Research Institute)

UN Global Pulse. (2017). *Using machine learning to analyse radio talk in Uganda.* Kampala: UN Global Pulse.

USAID. (2021). *Managing machine learning projects in international development: A practical guide.* Retrieved from USADI: https://www.usaid.gov/sites/default/files/documents/Vital_Wave_USAID-AIML-FieldGuide_FINAL_VERSION_1.pdf

van der Maaten, L., Postma, E., & van den Herik, J. (2009). Dimensionality Reduction: A Comparative Review. *J Mach Learn Res*. 10(66-71), 13.

White, H. (2019). The twenty-first century experimenting society: the four waves of the evidence revolution. *Palgrave Communications*. 5(1), 1-7.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D.,... Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11(1), 1-11.

Independent Evaluation Unit
Green Climate Fund
175 Art center-daero, Yeonsu-gu
Incheon 22004, Republic of Korea
Tel. (+82) 032-458-6450
ieu@gcfund.org
https://ieu.greenclimate.fund