Can machine learning algorithms predict publication outcomes? A case study of COVID-19 preprints

Sai Koneru
Information Sciences and Technology
The Pennsylvania State University
State College, PA, USA
sdk96@psu.edu

Xin Wei, Jian Wu
Department of Computer Science
Old Dominion University
Norfolk, VA, USA
{xwei001,j1wu}@odu.edu

Sarah Rajtmajer
Information Sciences and Technology
The Pennsylvania State University
State College, PA, USA
smr48@psu.edu

***Abstract*—The COVID-19 pandemic catalyzed a large body of scientific work, much of which was completed and disseminated with groundbreaking speed. A significant portion of COVID-related work was posted to preprint servers and COVID-related preprints were more widely cited than their counterparts. This work leverages information retrieval, natural language processing, and supervised learning to predict the subsequent publication, within a year, of COVID-related papers posted to preprint servers in peer-reviewed venues. Our work is inspired by prior work surveying human experts for the same task. We compare the performance of ML and human predictions and discuss the implications of our findings for scientific publishing. The findings demonstrate that the Multi-Layer Perceptron yielded the highest performance, achieving a macro F1 score of 0.674 on the held-out set. This underscores the challenge of accurately predicting the outcomes of the human peer review process. The data and code are available at https://github.com/Sai90000/preprint_prediction.git.**

***Index Terms*—machine learning, publication forecasting, preprints, science of science**

I. Introduction

Changes to research publishing were amongst the countless impacts of the COVID-19 pandemic [1], [2]. Many journals significantly shortened review and publication timelines [1], [3]; a study of COVID-19 related papers published on PubMed during early months of the pandemic found the median time between paper submission and acceptance to be 6 days [3]. The rapid turn-around of COVID-related work came in tandem with an increase in sheer volume of papers published – both COVID-related and otherwise [2]. Preprints became a significant source of information for the community. Of 125,000 COVID-19–related scientific articles published within 10 months of the first confirmed case, more than 30,000 were hosted by preprint servers [4].

Amid the flurry of COVID-related papers, long-standing discussions about the role of peer review processes have been revisited in the context of the global health crisis [5]. A study of the COVID "paperdemic" found that a substantial fraction of low-quality research papers including flawed and/or questionable data and methods appeared as preprints [6]. While, another study suggests that COVID-related publications were overrepresented in the Retraction Watch Database in 2020 [5]. Low-quality research poses real risks to broader scientific and public discourse, and these risks were magnified during the pandemic as discussion moved to social media and scientific misinformation and disinformation spread quickly.

One assessment that may be useful for the evaluation of a given preprint is a measure of whether the paper would be likely to pass peer-review processes. Recent work using crowd-based forecasting has demonstrated some success predicting whether COVID-related preprints would be published in a peer-reviewed venue within 1 year and, if so, whether the venue would be one with high impact factor [7]. However, collecting such predictions from groups of human experts is generally time-consuming and incentives are usually required.

Machine learning (ML) methods have shown comparable or even superior performance to humans on many complex tasks involving high-dimensional data such as object recognition [8], question-answering [9]. In many cases, e.g., deep neural networks, the performance of ML algorithms is attributed to large volumes of human- or automatically-labeled training data. These models can be easily overfitted with limited training data. They also typically lack explainability, which limits potential application and raises critical questions about bias and trust. Recent research has demonstrated that traditional machine learning models, e.g., Support Vector Machines (SVM), can match human performance on multi-class legal text labeling tasks [10]. Prior work has shown the capability of simple ML models for the task of predicting the outcomes of replication studies of published findings [11] – a task which has been studied extensively with human participants and which is similar to the task of publication prediction which is our focus here. In the context of this work, we use the terms *prediction* and *prediction within a year of posting on a preprint server* interchangeably, and they carry the same meaning.

Our study addresses three primary research questions.

- Can ML algorithms be used to predict the subsequent publication of a posted preprint in a peer-reviewed venue?
- Can ML algorithms match or outperform human experts on the task of predicting the publication of preprints?
- Which features of a preprint or its corresponding meta-data are strong predictors of subsequent publication in a peer-reviewed venue?

Following, we explore two widely used ML algorithms, Random Forest, Multi-Layer Perceptron, for publication pre-

diction. We compare the performance of these models to human expert predictions on the same dataset and discuss implications of these findings for scientific publishing.

II. Related work

Our work is built on an existing framework for content-based feature extraction from scientific papers [12]. This framework extracts 41 distinct features within five major categories from a scientific paper. Although developed for papers in the social and behavioral sciences, most feature extractors can be directly transferred to papers in other domains, e.g., acknowledgments, p-value expressions, because relevant patterns are standardized across the literature. The subset of features related to citation counts or publication venues were excluded from consideration for this work, as their inclusion would result in data leakage. Other frameworks have been proposed to extract scientific measurements from text, such as temperature sensor values in the geo-spatial context [13], but this framework is not applicable to most papers related to COVID-19.

To our knowledge, there are only a few papers that have attempted to predict the publication status of preprints. In recent work [14], after exploring 100 matched preprint–journal-article pairs using the NIH's iSearch COVID-19 Portfolio, authors showed that there was no obvious difference in expert opinions of preprints published vs. those not published. Using a simple regression, the authors attempted to predict the publication status of these papers but did not find a strong correlation between the publication status and the review scores of preprints. Another study [5] showed that COVID-related preprints are more likely to be published after peer reviews than non-COVID-19-related papers with a considerably shortened submission-to-acceptance time.

The closely-related task of citation prediction seeks to understand features that effectively predict citation counts or other impact factors. Citation prediction was used as a KDD cup challenge back in 2003 [15]; the winning team treated the task as a time-series prediction problem. Citation prediction typically involves using features derived from different aspects of a publication as inputs. Among these are the features from semantic information contained within the publication text. Studies have shown that the text complexity and sentiment of abstracts are positively correlated with future citations [16]. Analysis of publications from *American Economic Review* revealed that publications that are difficult to read tended to receive lower citations [17]. In contrast, the assessment of top-cited publications in neuro-imaging found that articles published in high-impact journals are less readable [18]. Another set of features are related to references within a publication; a study of publications from the Nanoscience and Nanotechnology fields showed that reference counts and their impact are most effective determinants of future citation counts [19]. Reference counts were shown to be strong predictors of citations in other fields as well [20]. Another feature with predictive power is acknowledgement of funding; a paper's funding status has a statistically significant positive correlation to the number of citations it receives [21]. Studies show that collaborative articles received higher visibility [22]. Furthermore, author prestige impacts peer review decisions [23]. Statistical information within a publication also impacts the peer review studies analyzing published literature shows that p-values reported effect the peer review decisions [24]. In another paper [25], the authors proposed a model based on paper metadata to predict the long-term citation count. A recent paper [26] provides a comprehensive review of the citation prediction literature and proposes a taxonomy consisting of six types of features and four types of forecasting methods.

The social impact of a publication has also been studied. Studies have shown a stronger correlation between citation-based metrics and the quality of a publication than the number of tweets [27]. Furthermore,a recent study that tracks changes between preprints found that although preprints undergoing discrete textual changes are commented upon and cited more often, the amount of attention given to preprints does not reflect their impact upon publication [28].

Our work is the first to investigate machine learning to predict publication status using a comprehensive list of content-based features.

III. Data

In this section, we provide an overview of the data reduction process from the raw data to the final data. We elaborate details of each step in the process in subsequent sections.

This study considers the set of preprint articles curated for a prior study surveying researchers to predict the eventual publication of preprints in peer-reviewed outlets [7]. The original dataset contains 8,176 preprints, collected from the bioRxiv and medRxiv *COVID-19 SARS-CoV-2 collection*, uploaded to the repositories between January 2020 and August 2020. We collected their full text PDFs using the repositories' application programming interfaces (APIs). PDFs were input to a feature extraction pipeline developed for previous work [11] and further detailed below. Certain features have been shown important for reproducibility assessments [11].

After the feature extraction pipeline, the dataset was reduced to 5,893 preprints. There are several reasons that could lead to a reduction of sample size. First, in the feature extraction step, a fraction of PDFs cannot be successfully converted to text. Therefore, the DOIs may not be able to be correctly extracted from the full text. The second reason was that the papers were not indexed by Semantic Scholar [29] (SS), so certain features (e.g., reference features) could not be obtained. After feature extraction, we built the ground truth by identifying preprints that were officially accepted for publication. The workflow that shows the data reduction process is illustrated in Fig. 1.

IV. Feature Extraction

A. Content-based Features

An automated pipeline generates features from each paper's full text as an input. The PDFs were first converted to text using GROBID. The pipeline outputs up to 18 features for each paper in 4 categories: *Reference features*, *Authorship features*,

Statistical features and the *Acknowledgement feature* (Table I). Because our goal is to predict publication of preprint, we focus on the pre-publication features, excluding post-publication features like citation counts.

- *Reference features.* This group contains 5 features extracted from preprint's references, obtained from SS API. It includes number of important references and references that are cited for different purposes, such as background, methodology, and results.
- *Authorship features.* This group contains 2 shallow features: the number of authors and the university rank of the leading author. We collected university ranking data from the 2020 Times Higher Education rankings and use it as a lookup table to determine the feature value.
- *Statistical features.* This group contains 10 features focusing on extracting p-values, reported associated with hypothesis tests. The extraction method was adopted from [30], which uses regular expressions against the full text to identify 10 most frequently used tests, such as t-test, and f-test. This method also aggregates multiple p-values when found in a preprint. Additionally, we calculate readability, sentiment, and subjectivity of abstracts. Readibility is calculated using the Flesch–Kincaid score [31]. Sentiment is calculated using the AllenNLP package based on the RoBERTa_Large model [32]. The subjectivity is calculated using the TextBlob package [33].
- *Acknowledgement feature.* The acknowledgment feature is a binary feature indicating whether a funding agency is acknowledged. We adopt the hybrid method proposed in [34], which achieved a performance of $F1 = 0.92$.

It should be noted that for a given preprint, not all features may be available. In this case, a null value distinct from all possible values is used.

B. Early Popularity Features

We acquired eight features that are early indicators of preprint popularity by automatically scraping data from the Altmetric pages for the papers in our corpus, e.g., https://biorxiv.altmetric.com/. These features can be divided into two different groups.

- *Usage features.* Features like first-month PDF downloads and abstract views, are valuable for gauging community engagement with preprints. These metrics are good indicators of publication status for papers submitted to high-impact journals [35]. Features of this type include the first month PDF downloads and the rate, the first-month abstract views and the rate, and the first-month full text views and the rate. The bioRxiv and medRxiv servers track and update the number of abstracts, full-text HTML views, and PDF downloads on a daily basis and aggregate them monthly. This means that the daily counts for usage metrics are not accessible during our data collection. This can potentially, introduce a bias, as preprints uploaded at the start of a calendar month, which can garner higher feature values, with respect to the values obtained when the preprints were submitted at the end of the month. To address this, we added the change rates of these features by calculating the difference of counts between the first and the second calendar months as features.
- *Social interaction features.* Features that capture how the contents create community engagement through social interactions are shown to correlated with early citations [36]. Features of this type include the number of tweets within a month, and the Altmetric score [37]. We counted number of tweets within the first 30 days using the tweet timestamps relative to preprint upload. To quantify the level of attention received by a preprint, we use its initial Altmetric score [38]. An Altmetric score is a weighted count of various indicators such as the number of news articles and the number of blog mentions. The features were obtained from the data from the prior work [7].

C. Result Type Features

We extracted a categorical feature representing the type of outcomes of the study presented within a preprint, using the information provided by the reprint repositories. These were provided by the authors during the process of submitting preprint to the server and can be one of the following class.

- *New* results bring a new perspective and advance the field;
- *Contradictory* results of replication studies either fail to support or contradict the findings of prior published research;
- *Confirmatory* results of replication studies validate and offer support to the findings in previous publications;
- *Neutral* results do not fall into any of the aforementioned categories.

In the final dataset (Fig. 1) of 5302 samples (2613+2689), a majority (78.5%) of the preprints within the full dataset fall into the *neutral* category whereas 21.0% of preprint results are identified as *new*. The *contradictory* and *confirmatory* results comprise less than half percent of the preprints.

V. Curation of Ground Truth

Ground truth curation involved obtaining corresponding peer-reviewed publication dates for each preprint. This was achieved by integrating information from the preprint repositories as well as scholarly databases specifically SS and CrossRef [39] (CR).

The process had two stages. In the first stage, published DOIs were collected for the preprints identified as published by the Rxiv repositories (bioRxiv and medRxiv). The publication dates for these DOIs were then queried from SS and CR. In the second stage, preprints for which the repositories did not explicitly mention a peer-reviewed publication were searched in the SS and CR databases using their titles as a queries. The aim of this stage was to overcome the repository coverage limitation. Matching between preprints and search results in the scholarly databases was performed based on title, abstract, and authorship [40]. The relevance was determined

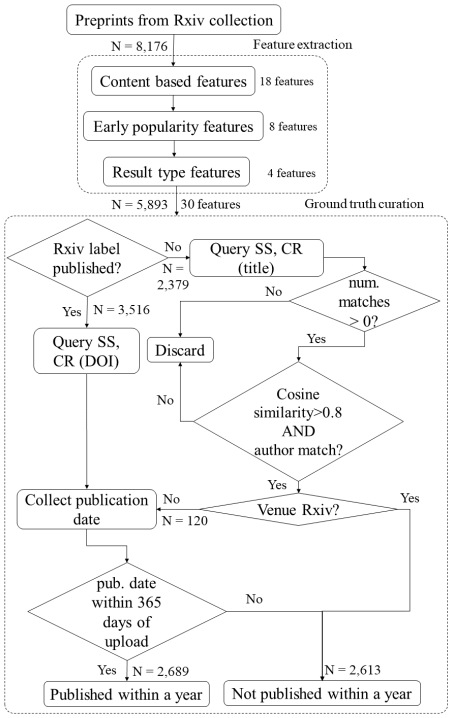

Fig. 1. Flowchart showing different steps involved in feature extraction and ground truth curation.

using document-level embeddings, from concatenated title and abstract, by pre-trained SPECTER model [41]. A cosine similarity threshold of 0.8, and a complete in author lists to identify positive matches. Positive matches with venues other than Rxiv, were identified as peer-reviewed publications and their publication dates were obtained from SS and CR APIs. Preprint searches resulting in single positive matches with Rxiv venues were labeled as *not published*. The second stage resulted in matching 120 preprints to their peer-reviewed publications.

The difference between the date of preprint upload and the corresponding publication date of the match was then used to determine if a preprint was peer-reviewed and published within a year of upload. Preprints published in peer-reviewed venues less than 365 days of upload to the Rxiv servers were labeled as *published within a year*. Preprints that took more than a year to publish were labeled as *not published within a year*. Additionally, preprints that resulted in a single Rxiv match (only the preprint itself) during the second stage of the matching process were also labeled as *not published within a year*. Finally, preprints without a *published* label from the repositories which also did not yield any matches from SS and CR search were discarded as their publication status could not be established.

This process resulted in a dataset with a total of 5,302 preprints and a well-balanced class distribution having 50.7% of preprints labeled as *published* (published within a year), and the remaining 49.3% with a label *not published* (within a year).

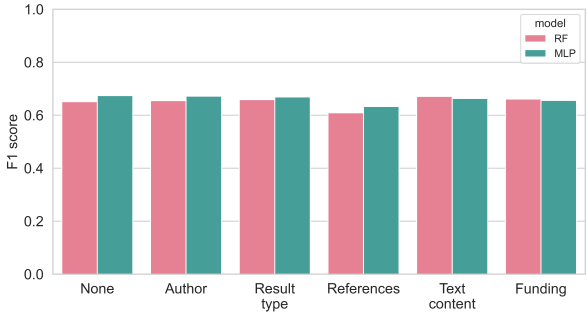

Fig. 2. Summary of model performance on held-out data (*N=1012*) when features of a particular type are removed.

VI. Machine Learning-Based Forecasting Model

A. Pre-processing

We binarized the prediction lable such that 1 represents *published within a year* label while 0 represents *not published within a year*. To prepare the features for modeling we employed two widely used preprocessing techniques

a) One hot encoding: We implemented one-hot encoding technique to non-ordinal categorical variable*result type* which has 4 distinct classes. This resulted in creation of four new binary features one for each class.

b) Feature scaling: We standardized the numerical features to mitigate a potential unequal influence due to feature magnitudes. Each feature was transformed to have a mean of zero and a standard deviation of one.

The feature extraction and preprocessing steps resulted in a total of 30 different features from 6 mutually exclusive groups. The complete list of extracted features and their short descriptions are provided in the Table I. Among the 5,302 preprints, 242 were in the dataset from the study involving surveys with human participants asked to predict publication [7]. These 242 were part of the original 400 preprints selected based on their relatively high Altmetric scores. These specific preprints were isolated for the purpose of performing a comparative analysis comparing the performance of machine learning models to human participants. From the remaining data, we randomly chose 80% (N = 4048) for training and kept the remaining 20% (N = 1012) as held-out test data.

B. Modeling

We model this problem as a binary classification task with the class labels *published* (published within a year), *not published* (-within a year). We use two different supervised machine learning models to predict the publication status of a preprint using the features discussed in the previous subsections. Specifically, we trained Random Forest (RF) using Gini impurity index to measure the quality of a split and Multilayer Perceptron (MLP) which can model non-linear relationships.

C. Parameter tuning

To prevent model overfitting, we tuned their hyperparameters using Optuna framework [42]. The approach involved

TABLE I
THE PHI_K, POINT BISERIAL CORRELATIONS, AND CORRESPONDING SIGNIFICANCE VALUES, BETWEEN THE PREDICTION LABEL AND CONTINUOUS, CATEGORICAL FEATURES RESPECTIVELY.

Feature	Category	Brief description	Correlation	*P*
influential reference count	Reference	number of references which had a strong impact on the citing paper	-0.063	0.000
reference background		reference intent classified as background	-0.047	0.000
reference methodology		reference intent classified as methodology	-0.076	0.000
reference result		reference intent classified as results	-0.089	0.000
references count		number of references the target paper cites	0.147	0.000
author count	Authorship	total number of authors	0.130	0.000
university rank		normalized score based on 2020 Times Higher Education rankings	-0.054	0.000
reading score	Statistical	Flesch–Kincaid score of the abstract	-0.232	0.000
subjectivity		Extent of personal opinion rather than factual information of abstract	-0.009	0.503
n_hypothesis tested		number of statistical tests found	0.053	0.000
real_p		minimum p-value among all the p-values extracted	0.014	0.307
real_p_sign		sign for the p-value ($=$, $<$, or $>$)	0.047*	0.000
p value range		the difference of the highest and the lowest p-value	0.063	0.000
extended p		whether the p-value features are associated with a test	0.017*	0.171
sentiment		extent of positive vs. negative of the abstract	0.053*	0.011
sample size		number of observations in the experiment data	-0.019	0.148
n_significant		total number of significant p-values (≤ 0.05)	0.061	0.000
funding status	Acknowledgment	exists acknowledgement of a funding agency	0.192*	0.000
tweets within a month	Popularity	number of tweets referencing the article within first 30 days of upload	0.037	0.006
first month download.		PDF downloads within the first calendar month of submission	0.036	0.008
first month abstract views		number of abstract views during the first calendar month of submission	0.042	0.002
first month full text views		variable corresponding to counts of full-text HTML views	0.018	0.183
abstract views rate		change of abstract views from first to second months of submission	-0.037	0.007
PDF dwnld. rate		rate of change of PDF downloads	-0.033	0.016
rate of full text views		rate of change in counts of full-text HTML views	-0.009	0.477
altmetric score		a measure of research impact through online interactions	0.059	0.000
new results	Result type	indicator corresponding to an advance in a field	0.106*	0.000
confirmatory results		studies with results that replicate/confirm a previously work	0.0*	0.658
contradictory results		studies that contradict/fail to replicate results from prior work	0.0*	1.0
neutral results		general category for findings that do not fit any other *result types*	0.104*	0.000

* *Phi_k* correlation.

Bayesian hyperparameter tuning with the objective of maximizing 5-fold cross-validation F1 score on training data. Each tuning experiment consisted of a total of 1500 trials with hyperparameters with numeric values sampled from a uniform distribution over their respective minimum and maximum values. The categorical hyperparameters were sampled using a multinomial distribution with equal probability for each category. Table II shows different parameters for each of the models as well as their value ranges from which they were sampled. For RF model, we tuned the number of trees, the maximum tree depth, and the minimum samples per leaf node. For the MLP model, hidden layer sizes and corresponding activation functions, learning rate, and weights optimization solver were tuned.

VII. RESULTS

In this section we summarize the results from our experiments on predicting preprint publication within a year of appearing on arXiv repository. Both tested models achieved moderate macro F1 scores on the held-out data (RF: 0.674, MLP: 0.651).

A. Correlation analysis

We sought to understand the association between the features and prediction labels. We used the Point Biserial correlation coefficient [43] to quantify the relationship between the features with continuous values and the dichotomous prediction label. To measure the correlation between binary features and the prediction label, phi_k correlation coefficient was used [44]. The correlation values and corresponding significance values are summarized in Table II. Features *reference count, author count* representing the number of references a preprint has and the number of authors respectively, are weakly but positively correlated with publication labels at high significance indicating a moderate relationship. Similarly, the prediction label is weakly correlated to binary features *funding status, new results, neutral results*. On the other hand, *reading score* showed a weak inverse correlation. Interestingly, the magnitude of the correlation for other features, none of them are strongly correlated with the prediction label.

B. Ablation study

To assess the contribution of each feature group to the prediction, we conducted ablation studies. We systematically removed each feature group and trained RF, and MLP models with the rest of the features. Hyperparameter tuning was done using the approach discussed in *Parameter Tuning* subsection. The findings, shown in Fig. 2, highlight the reliance of both the models on features derived from a preprint's references. This may be due to the reference features acting as a proxy for how well the contents of a preprint are grounded and justified. Conversely, removing other feature groups had minimal

TABLE II
SEARCH SPACE USED FOR TUNING EACH HYPERPARAMETER.

Model	Parameter	Range
RF	Number of estimators	[3,110]
	Maximum depth of trees	[3,30]
	Minimum samples in leaf	[10,100]
MLP	Initial learning rate	[0.001,0.1]
	Activation function	[logistic, tanh, ReLU]
	Optimization solver	[ADAM, SGD, LBFGS]
	Hidden layer 1 size	[5,25]
	Hidden layer 2 size	[5,25]

impact on performance, aligning with the observations from correlation analysis. However, model performance robustness suggests that models capture complex relationships between the features while compensating for the absence of a specific subset of features. Furthermore, these results show that despite low correlation values, the ML models have a decent performance. This could be due to some complex nonlinear relations between variables and the prediction label.

C. Feature importance

We used model agnostic permutation feature importance technique to assess feature relevance in model predictions. This involved evaluating model performance by randomly shuffling one feature at a time, while leaving other features unchanged, breaking its relationship with the prediction label. This is repeated for a set number of iterations and the significance of each feature is measured by the degree of model performance decrease due to shuffling, with a larger drop indicating greater model dependency on the feature.

Figs 3, 4 show permutation importance for baseline RF and MLP models respectively. The models were trained using all the features, and the permutation feature importance was calculated on the held-out test data by randomly shuffling each feature. Although the order of importance for both the models is same, the RF model exhibits a lower magnitude of performance drop compared to the MLP model. Furthermore, the spread in drop for most features ranges from negative to positive values. This inconsistency in performance drop indicates that the RF model does not heavily rely on specific individual features but instead leverages a combination of features for prediction. This aligns with the findings from the ablation study where the model performance remained relatively stable even after the removal of features from certain feature groups. Additionally, features *Altmetric score, full text views in the first month, rate of change in full text views* had a negative impact on prediction performance.

Conversely, the importance plot for the MLP model (Fig. 4) shows a consistent positive drop in performance for features *University Rank, Number of tweets within first month, Subjectivity, Sentiment, Sample size, and Number of references* and a consistent lack of impact of features created from *result type*.

D. Comparison with forecasts by human participants

The human participants forecasted publication status of a total of 400 preprints of which, as mentioned earlier, we were able to extract features and establish ground truth for 242 preprints. We used this subset as a held-out set and to compare performance of human forecasters and the trained algorithmic models. We compared the accuracy scores for the forecasts made by humans to the machine learning models' predictions. Humans predictions have an accuracy score of 57.0% with an macro F1 score of 0.518 which is better than a random chance. Both the baseline RF and MLP models, with macro F1 scores of 0.599 and 0.574 respectively, performed better than human participants. Further analysis of the overlap between the predictions and the forecasts among the 242 preprints, human participants accurately forecasted the publication status for 138 preprints while missing on remaining 104. Among these 104, the base RF model correctly predicted the publication status of 89, where as base MLP model predicted 90 correctly. The base RF model made 79 incorrect predictions of which human participants correctly predicted 64. Similarly, human participants correctly predicted the outcomes for 66 of 80 preprints where MLP were incorrect. These differences become evident when considering precision and recall. Human participants demonstrated high precision (0.86) but low recall (0.37) for identifying the class *published* (within a year) whereas the models exhibited moderate precision (RF: 0.68; MLP: 0.67) and high recall (RF: 0.93; MLP: 0.93). Additionally, human participants had a lower precision score (0.46) compared to the models (RF: 0.72, MLP: 0.68) for *not published* (within a year) class. Although the sample size is smaller, this highlights that there are certain groups of preprints where human participants made better predictions than algorithmic models and vise-versa.

VIII. DISCUSSION

The task of predicting preprint publication outcomes is incredibly challenging because it fundamentally assumes there are reliable patterns in peer review that can be captured computationally. Yet, calls are emerging from researchers and journals alike for major overhauls to review and publication processes, precisely because they are too often incomplete, inconsistent, and biased [45]. As a community we currently equate success in peer review with research quality and reliability, whereas looking ahead we might focus on developing ML and AI to directly measure research quality, i.e., through replication prediction or generalizability assessment. Until then, arguably the most important contribution of any success in automated prediction of peer review outcomes is quantitative insight, and ideally explanation, of the current publishing processes including its many flaws.

Given the caveats around publication prediction we have just mentioned, our study also carries some more specific limitations both conceptual and technical. One conceptual assumption is that the features we extract from preprints and metadata contain sufficient signal for the task. Ultimately, our features are the intersection of the set of features we expect may have relevance given established literature and the set of features we are technically able to extract. For example, in prior work we have surveyed researchers to ask what they

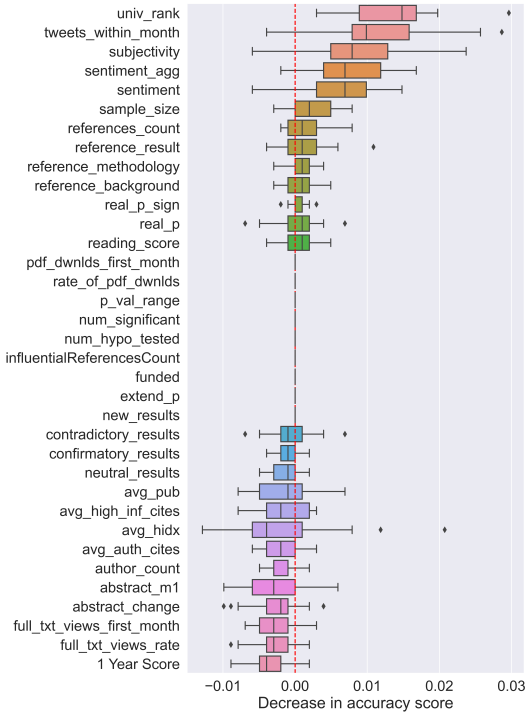

Fig. 3. Permutation feature importance scores on held-out data for RF model trained with all the features within the dataset.

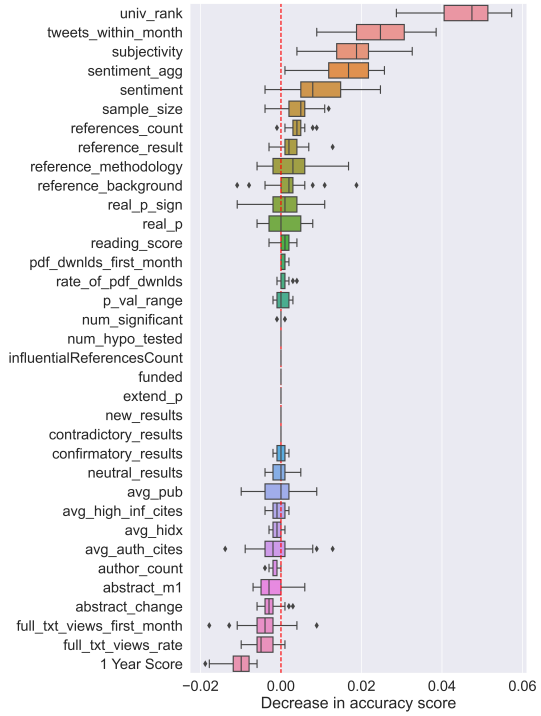

Fig. 4. Plot showing feature importance scores on held-out data and MLP model.

look for when evaluating of a published finding. Nuances of study design, basis in theory, and many other factors were mentioned by researchers but are not easily captured with current NLP and information extraction technologies. A second set of limitations centers around the messiness of the Rxiv records, incompleteness of information available through APIs, and data preprocessing. For example, our approach would not capture instances wherein a preprint was uploaded at or after the date of peer-reviewed publication. In addition, popularity features are constrained to abstract views as Altmetrics provides full text views aggregated monthly.

Finally, we highlight the uniqueness of the COVID-19 context that surrounded these preprints and consequently our work here. As we noted earlier, COVID-related preprints were meaningfully different that their non-COVID counterparts. They were more frequently cited, more widely discussed, and subsequently published elsewhere with greater probability and on a shorter timeline. Insights specific to this context are valuable, as we seek to derive lessons from the COVID-19 pandemic and ready for future crises. However, future work should test the performance of the features and ML models we describe here on non-COVID baselines to understand their generalizability in adjacent literatures.

IX. Conclusion

In this work we have explored the use of algorithmic machine learning models to predict subsequent peer-reviewed publication of COVID-19 preprints. Using automated methods, we extract 30 different features from preprint text and metadata, reprenting information falling broadly into 6 mutually exclusive dimensions. Through complimentary analyses, ablation study and permutation feature importance, we have observed that while ML models can capture complex non-linear relationships between the features, they struggle to achieve high performance in predicting publication outcomes.

Furthermore, we have compared our predictions to forecasts made by human participants on the same papers. Although ML models outperform human participants overall, complimentary capabilities are highlighted. That is, human participant-driven forecasts performed better where ML models had sub-optimal performance and vice versa. Our findings suggest an opportunity for hybrid approaches combining crowd-sourcing and ML models to achieve a *best of both worlds* scenario.

Acknowledgment

We acknowledge support by DARPA W911NF-19-2- 0272. This work does not necessarily reflect the position or policy of DARPA and no official endorsement should be inferred. We thank Dr. Thomas Pfeiffer and his team for sharing the dataset collected for their study.

References

[1] M. B. Eisen, A. Akhmanova, T. E. Behrens, and D. Weigel, "Publishing in the time of covid-19," p. e57162, 2020.
[2] H. Else, "Covid in papers," *Nature*, vol. 588, no. 24/31, p. 553, 2020.
[3] A. Palayew, O. Norgaard, K. Safreed-Harmon, T. H. Andersen, L. N. Rasmussen, and J. V. Lazarus, "Pandemic publishing poses a new covid-19 challenge," *Nature Human Behaviour*, vol. 4, no. 7, pp. 666–669, 2020.

[4] N. Fraser, L. Brierley, G. Dey, J. K. Polka, M. Pálfy, F. Nanni, and J. A. Coates, "The evolving role of preprints in the dissemination of covid-19 research and their impact on the science communication landscape," *PLoS biology*, vol. 19, no. 4, p. e3000959, 2021.

[5] I. Kodvanj, J. Homolak, D. Virag, and V. Trkulja, "Publishing of covid-19 preprints in peer-reviewed journals, preprinting trends, public discussion and quality issues," *Scientometrics*, vol. 127, no. 3, pp. 1339–1352, 2022.

[6] R. J. Dinis-Oliveira, "COVID-19 Research: Pandemic Versus "Paperdemic", Integrity, Values and Risks of the "Speed Science"," *Forensic Sciences Research*, vol. 5, no. 2, pp. 174–187, 06 2020.

[7] M. Gordon, M. Bishop, Y. Chen, A. Dreber, B. Goldfedder, F. Holzmeister, M. Johannesson, Y. Liu, L. Tran, C. Twardy *et al.*, "Forecasting the publication and citation outcomes of covid-19 preprints," *Royal Society open science*, vol. 9, no. 9, p. 220440, 2022.

[8] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using yolo: challenges, architectural successors, datasets and applications," *Multimedia Tools and Applications*, vol. 82, no. 6, pp. 9243–9275, 2023.

[9] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty, "Building watson: An overview of the deepqa project," *AI Magazine*, vol. 31, no. 3, pp. 59–79, Jul. 2010.

[10] T. Orosz, R. Vági, G. M. Csányi, D. Nagy, I. Üveges, J. P. Vadász, and A. Megyeri, "Evaluating human versus machine learning performance in a legaltech problem," *Applied Sciences*, vol. 12, no. 1, 2022.

[11] J. Wu, R. Nivargi, S. S. T. Lanka, A. M. Menon, S. A. Modukuri, N. Nakshatri, X. Wei, Z. Wang, J. Caverlee, S. M. Rajtmajer *et al.*, "Predicting the reproducibility of social and behavioral science papers using supervised learning models," *arXiv preprint arXiv:2104.04580*, 2021.

[12] S. Rajtmajer, C. Griffin, J. Wu, R. Fraleigh, L. Balaji, A. Squicciarini, A. Kwasnica, D. Pennock, M. McLaughlin, T. Fritton *et al.*, "A synthetic prediction market for estimating confidence in published work," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 13 218–13 220.

[13] M. A. Suryani, Y. Wölker, D. Sharma, C. Beth, K. Wallmann, and M. Renz, "A framework for extracting scientific measurements and geo-spatial information from scientific literature," in *2022 IEEE 18th International Conference on e-Science (e-Science)*, 2022, pp. 236–245.

[14] L. Nelson, H. Ye, A. Schwenn, S. Lee, S. Arabi, and B. I. Hutchins, "Robustness of evidence reported in preprints during peer review," *The Lancet Global Health*, vol. 10, no. 11, pp. e1684–e1687, 2022.

[15] J. Manjunatha, K. Sivaramakrishnan, R. K. Pandey, and M. N. Murthy, "Citation prediction using time series approach kdd cup 2003 (task 1)," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 2, pp. 152–153, 2003.

[16] A. Fronzetti Colladon, C. A. D'Angelo, and P. A. Gloor, "Predicting the future success of scientific publications through social network and semantic analysis," *Scientometrics*, vol. 124, pp. 357–377, 2020.

[17] B. C. McCannon, "Readability and research impact," *Economics Letters*, vol. 180, pp. 76–79, 2019.

[18] A. W. Yeung, T. K. Goto, and W. K. Leung, "Readability of the 100 most-cited neuroimaging papers assessed by common readability formulae," *Frontiers in human neuroscience*, vol. 12, p. 308, 2018.

[19] F. Didegah and M. Thelwall, "Determinants of research citation impact in nanoscience and nanotechnology," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 5, pp. 1055–1064, 2013.

[20] N. Onodera and F. Yoshikane, "Factors affecting citation rates of research articles," *Journal of the Association for Information Science and Technology*, vol. 66, no. 4, pp. 739–764, 2015.

[21] S. Roshani, M.-R. Bagherylooieh, M. Mosleh, and M. Coccia, "What is the relationship between research funding and citation-based performance? a comparative analysis between critical disciplines," *Scientometrics*, vol. 126, no. 9, pp. 7859–7874, 2021.

[22] V. Larivière, Y. Gingras, C. R. Sugimoto, and A. Tsou, "Team size matters: Collaboration and scientific impact since 1900," *Journal of the Association for Information Science and Technology*, vol. 66, no. 7, pp. 1323–1332, 2015.

[23] J. Huber, S. Inoua, R. Kerschbamer, C. König-Kersting, S. Palan, and V. L. Smith, "Nobel and novice: Author prominence affects peer review," *Proceedings of the National Academy of Sciences*, vol. 119, no. 41, p. e2205779119, 2022.

[24] S. Hopewell, K. Loudon, M. J. Clarke, A. D. Oxman, and K. Dickersin, "Publication bias in clinical trials due to statistical significance or direction of trial results," *Cochrane Database of Systematic Reviews*, no. 1, 2009.

[25] A. Ma, Y. Liu, X. Xu, and T. Dong, "A deep-learning based citation count prediction model with paper metadata semantic features," *Scientometrics*, vol. 126, no. 8, pp. 6803–6823, 2021.

[26] W. Xia, T. Li, and C. Li, "A review of scientific impact prediction: tasks, features and methods," *Scientometrics*, vol. 128, no. 1, pp. 543–585, 2023.

[27] L. Bornmann and R. Haunschild, "Do altmetrics correlate with the quality of papers? a large-scale empirical study based on f1000prime data," *PloS one*, vol. 13, no. 5, p. e0197133, 2018.

[28] L. Brierley, F. Nanni, J. K. Polka, G. Dey, M. Pálfy, N. Fraser, and J. A. Coates, "Tracking changes between preprint posting and journal publication during a pandemic," *PLoS biology*, vol. 20, no. 2, p. e3001285, 2022.

[29] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. S. Weld, "S2orc: The semantic scholar open research corpus," *arXiv preprint arXiv:1911.02782*, 2019.

[30] S. S. T. Lanka, S. M. Rajtmajer, J. Wu, and C. L. Giles, "Extraction and evaluation of statistical information from social and behavioral science papers," in *Companion Proceedings of the Web Conference 2021*, 2021, pp. 426–430.

[31] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Naval Technical Training Command Millington TN Research Branch, Tech. Rep., 1975.

[32] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[33] S. Loria *et al.*, "textblob documentation," *Release 0.15*, vol. 2, no. 8, 2018.

[34] J. Wu, P. Wang, X. Wei, S. M. Rajtmajer, C. L. Giles, and C. Griffin, "Acknowledgement entity recognition in CORD-19 papers," in *Proceedings of the First Workshop on Scholarly Document Processing, SDP@EMNLP 2020*. Association for Computational Linguistics, 2020, pp. 10–19.

[35] R. J. Abdill and R. Blekhman, "Tracking the popularity and outcomes of all biorxiv preprints," *Elife*, vol. 8, p. e45133, 2019.

[36] X. Shuai, A. Pepe, and J. Bollen, "How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations," *PloS one*, vol. 7, no. 11, p. e47523, 2012.

[37] J. Priem, D. Taraborelli, P. Groth, and C. Neylon, "Altmetrics: A manifesto," 2011.

[38] N. S. Trueger, B. Thoma, C. H. Hsu, D. Sullivan, L. Peters, and M. Lin, "The altmetric score: a new measure for article-level dissemination and impact," *Annals of emergency medicine*, vol. 66, no. 5, pp. 549–553, 2015.

[39] R. Lammey, "Crossref text and data mining services," *Insights*, vol. 28, no. 2, 2015.

[40] P. Eckmann and A. Bandrowski, "Preprintmatch: A tool for preprint to publication detection shows global inequities in scientific publication," *Plos one*, vol. 18, no. 3, p. e0281659, 2023.

[41] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld, "Specter: Document-level representation learning using citation-informed transformers," *arXiv preprint arXiv:2004.07180*, 2020.

[42] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[43] R. F. Tate, "Correlation between a discrete and a continuous variable. point-biserial correlation," *The Annals of mathematical statistics*, vol. 25, no. 3, pp. 603–607, 1954.

[44] M. Baak, R. Koopman, H. Snoek, and S. Klous, "A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics," *Computational Statistics & Data Analysis*, vol. 152, p. 107043, 2020.

[45] M. B. Eisen, A. Akhmanova, T. E. Behrens, J. Diedrichsen, D. M. Harper, M. D. Iordanova, D. Weigel, and M. Zaidi, "Peer review without gatekeeping," p. e83889, 2022.