Computers in Biology and Medicine 110 (2019) 29–39

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Efficient learning from big data for cancer risk modeling: A case study with melanoma

Aaron N. Richter[*], Taghi M. Khoshgoftaar

Department of Computer & Electrical Engineering and Computer Science College of Engineering and Computer Science, Florida Atlantic University, 777 Glades Road EE 403, Boca Raton, FL, 33431-0991, USA

ARTICLE INFO

Keywords:
Big data
Cloud computing
Machine learning
Electronic health records
Early detection of cancer

ABSTRACT

Background: Building cancer risk models from real-world data requires overcoming challenges in data preprocessing, efficient representation, and computational performance. We present a case study of a cloud-based approach to learning from de-identified electronic health record data and demonstrate its effectiveness for melanoma risk prediction.

Methods: We used a hybrid distributed and non-distributed approach to computing in the cloud: distributed processing with Apache Spark for data preprocessing and labeling, and non-distributed processing for machine learning model training with *scikit-learn*. Moreover, we explored the effects of sampling the training dataset to improve computational performance. Risk factors were evaluated using regression weights as well as tree SHAP values.

Results: Among 4,061,172 patients who did not have melanoma through the 2016 calendar year, 10,129 were diagnosed with melanoma within one year. A gradient-boosted classifier achieved the best predictive performance with cross-validation (AUC = 0.799, Sensitivity = 0.753, Specificity = 0.688). Compared to a model built on the original data, a dataset two orders of magnitude smaller could achieve statistically similar or better performance with less than 1% of the training time and cost.

Conclusions: We produced a model that can effectively predict melanoma risk for a diverse dermatology population in the U.S. by using hybrid computing infrastructure and data sampling. For this de-identified clinical dataset, sampling approaches significantly shortened the time for model building while retaining predictive accuracy, allowing for more rapid machine learning model experimentation on familiar computing machinery. A large number of risk factors (> 300) were required to produce the best model.

1. Introduction

In 2018, there were an estimated 1,735,350 new cases of cancer in the U.S [1]. Nevertheless, it is economically infeasible to screen over 320 million people in the U.S. for all types of cancer, and some cancers do not have a screening test or have not shown any improvements in detection from such a test. In this context, cancer risk models could provide immense value for informing screening guidelines by facilitating the identification and close follow-up of high-risk patients. In addition, they would enable more shared decision making between physicians and patients by providing evidence-based estimates of disease risk and prognosis [2].

A literature search retrieved several reports of predictive models for cancer risk [3]. We identified several shortcomings. *Availability of structured clinical data*: Structured data points regarding patient history and encounters are limited. Many data-capture systems record free-text notes that are difficult to standardize across several patient charts. Data sharing among healthcare providers is lacking, limiting holistic views of patient history. *Old data:* Most studies were published five or more years after the end of the study period. This results in stale models that might not reflect the current state of diagnosis and treatment. *Advanced modeling methods:* Researchers often only use one or two familiar algorithms, possibly because of a lack of experience with various tools or limitations in computing power.

In their review, Usher-Smith et al. found several case–control studies that compared melanoma populations to non-melanoma populations and built discriminators between them [4]. Like many cancers, melanoma screening is important given the poor survival rate of late-stage patients (five-year survival rate: 20% in patients with distant metastases vs. > 99% in early-stage patients). Moreover, the incidence

[*] Corresponding author.
E-mail addresses: arichter@fau.edu (A.N. Richter), khoshgof@fau.edu (T.M. Khoshgoftaar).

https://doi.org/10.1016/j.compbiomed.2019.04.039
Received 7 March 2019; Received in revised form 24 April 2019; Accepted 29 April 2019

of melanoma is growing, with 96,000 new cases expected in 2019 [5]. Thus, a risk prediction model could flag high-risk patients and enroll them in screening programs to detect melanoma early. In addition to learning from tabular data, there have been recent advancements in computer vision models to detect cancerous lesions from skin images [6,7].

Electronic health record (EHR) systems have been rapidly adopted in the U.S. over the last decade, largely because of requirements introduced by the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 [8]. In addition, the 21st Century Cures Act provided funding for increasing interoperability among EHR systems [9]. While there are structured coding vocabularies for some clinical information, many systems collect clinical narrative and procedure notes through dictation or free typing. Hence, the complexity of clinical information and differences among EHR systems make it difficult to build natural language-processing systems that synthesize structured data across multiple practices or hospitals [10].

If a large quantity of consistent patient data can be collected for a predictive model, computational challenges arise when transforming the data and training a machine learning algorithm. First, data elements must be extracted from the EHR system and transformed into a tabular format to be passed to a machine learning model. The size of the dataset and complexity of the machine learning algorithm can subsequently introduce computational challenges. The cloud, computing infrastructure accessed through the internet, enables users to launch machines of varying size with prebuilt libraries for machine learning algorithms. This technology can be utilized to evaluate a wide range of algorithms to produce the most accurate model. When dealing with big data, or data that cannot be processed through traditional architectures, predictive accuracy is not the only consideration when choosing classifiers and machine learning techniques; computational complexity and cost must also be factored in the selection process.

Here, we present a cloud-based approach to learning from big data and demonstrate its effectiveness on melanoma risk prediction from EHR system data. We evaluated methods for practical cost savings while maintaining model accuracy by using various types of computing infrastructures and data sampling techniques. Clinical utility of the models was assessed by examining the selected features and their impact on predicting melanoma risk for this population.

2. Materials & methods

2.1. Data

We used the Modernizing Analytics for Melanoma (MAMEL) dataset for the experiments [11]. The data were collected from a mobile-first, structured data-input, and cloud-based dermatology-specific EHR system and de-identified in accordance with HIPAA [12]. De-identified data were available from over 100 million dermatology visits throughout the U.S. recorded from January 1, 2011 to December 31, 2017.

The models in this study were built to predict melanoma diagnosis within 12 months of a given patient encounter. We included patients in the experiments if they had no evidence of melanoma (defined as ICD9 V10.82, ICD9 172.*, ICD10 Z85.820, ICD10 C43.*, ICD10 D03.*, melanoma SNOMED, biopsy result, or cancer log entry) through 2016. We then tracked the patients through 2017 to determine if they received a diagnosis of melanoma. This served as the binary class label for the prediction problem: “melanoma” or “no melanoma.” The visit from which predictions were made (the “index visit”) was selected based on the following criteria for positive and negative cases.

Positive cases: visit at least 6 months, and at most 12 months, before the earliest melanoma diagnosis in 2017. The earliest visit meeting this criterion was selected as the index visit. *Negative cases*: visit at least 12 months, and at most 24 months, before any visit in 2017. The latest visit meeting this criterion was selected as the index visit.

These constraints were inspired by Avati et al's approach for a hospital mortality prediction problem [13]. The goal of the constraints is to provide a consistent window of prediction time for positive and negative cases. The 6-month lower-bound for positive cases was selected to ensure the index visit was not a presentation for a melanoma biopsy or excision, and is a large enough time window to enact a change in screening patterns for early cancer detection. Furthermore, the lower bound of follow-up time for the negative cases must be greater than the upper bound of follow-up time for the positive cases. This is to ensure that the negative cases truly did not develop melanoma within a year. Otherwise, they may have developed it at some point after the observation time. We selected the 24-month upper-bound for negative cases to ensure that patients were consistently following up with their dermatologist. Patients that did not have an index visit matching these criteria were excluded from the study. We aggregated data from visits and prescriptions through each patient's index visit to use as independent variables.

For each patient, we collected three types of data from the MAMEL dataset: Patient Data, Visit Data, and Historical Visit Data (Appendix A). Patient Data refers to static patient data that is not collected longitudinally, such as age, sex, race, melanoma family history, geographic location (i.e., U.S. state), family history conditions, and drug allergies. Visit Data represents the data recorded in the patient encounter of the index visit: chief complaints, review of systems, vitals, skin exams, diagnoses, procedures, body locations evaluated, prescriptions, biopsy results, and medical codes generated from the visit. Historical Visit Data contains the same elements as the Visit Data category, but the features are aggregated across all visits prior to the index visit. As most data elements are categorical in nature, we aggregated each by counting the number of occurrences of each feature value across all visits. For the few numeric variables, we calculated summary statistics of each feature across the visits (minimum, maximum, mean, median, standard deviation). Appendix A describes each data element with a description, number of categories for each feature, and percentages of missing data. Interested readers can refer to our previous work for more background information and exploratory analysis of features [11]. The missingness of each variable in a random sample of patients is given in Fig. 1. Most Patient Data elements are required, and a visit generally contains complete data about the exam, diagnosis, procedure, and related ICD and CPT codes. There is more missing data for the Historical Visit Data, as not all patients had visits recorded before their index visit.

2.2. Preprocessing & sampling

We performed all data processing and model building in the cloud using Amazon Web Services (AWS).[1] We used AWS Hadoop clusters (Amazon Elastic MapReduce)[2] with the Apache Spark [14] processing engine to extract and process the de-identified data. We saved the final data for model training in a sparse matrix. We utilized sparse matrices to greatly compress the size of the data files, which allowed us to use libraries built for single-machine computing to perform the machine learning experiments. We used instances from the Amazon Elastic Compute Cloud (EC2)[3] to train machine learning models. EC2 supports many types of machines for different workloads with various numbers of CPU cores and amounts of memory. We explored the running time and cost of these instances for the different machine learning workloads.

The data extraction and matrix creation process is outlined in Fig. 2. Given that the Patient Data variables are not longitudinal, these data were extracted separately from longitudinal data such as patient encounters (or visits) and prescriptions. The longitudinal data were

[1] https://aws.amazon.com/.

[2] https://aws.amazon.com/emr/.

[3] https://aws.amazon.com/ec2/.

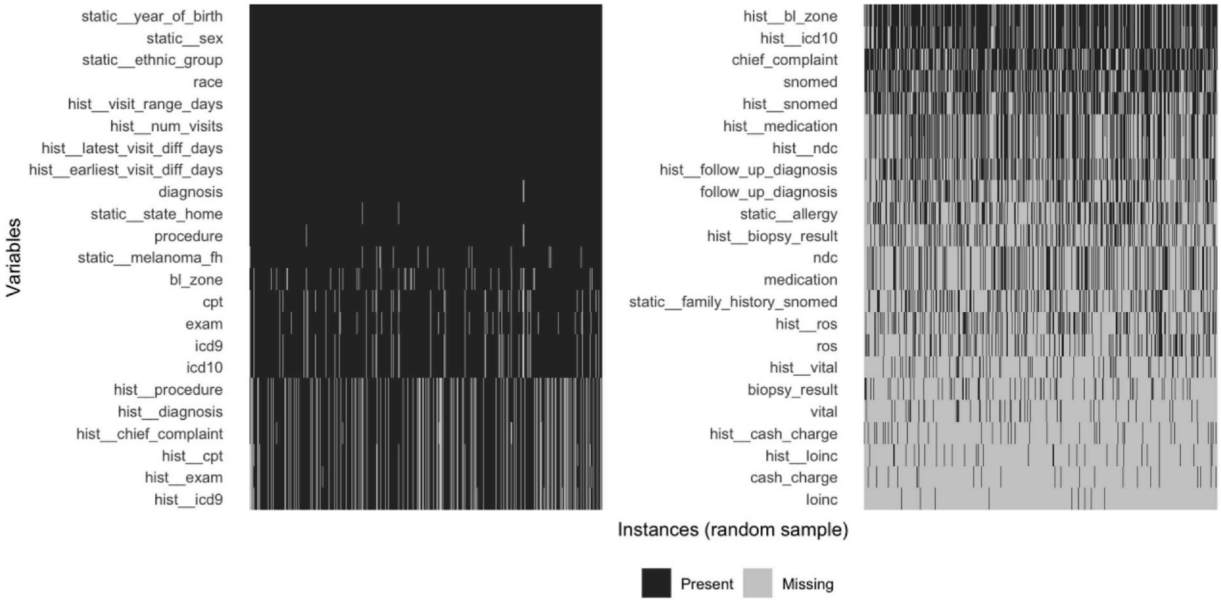
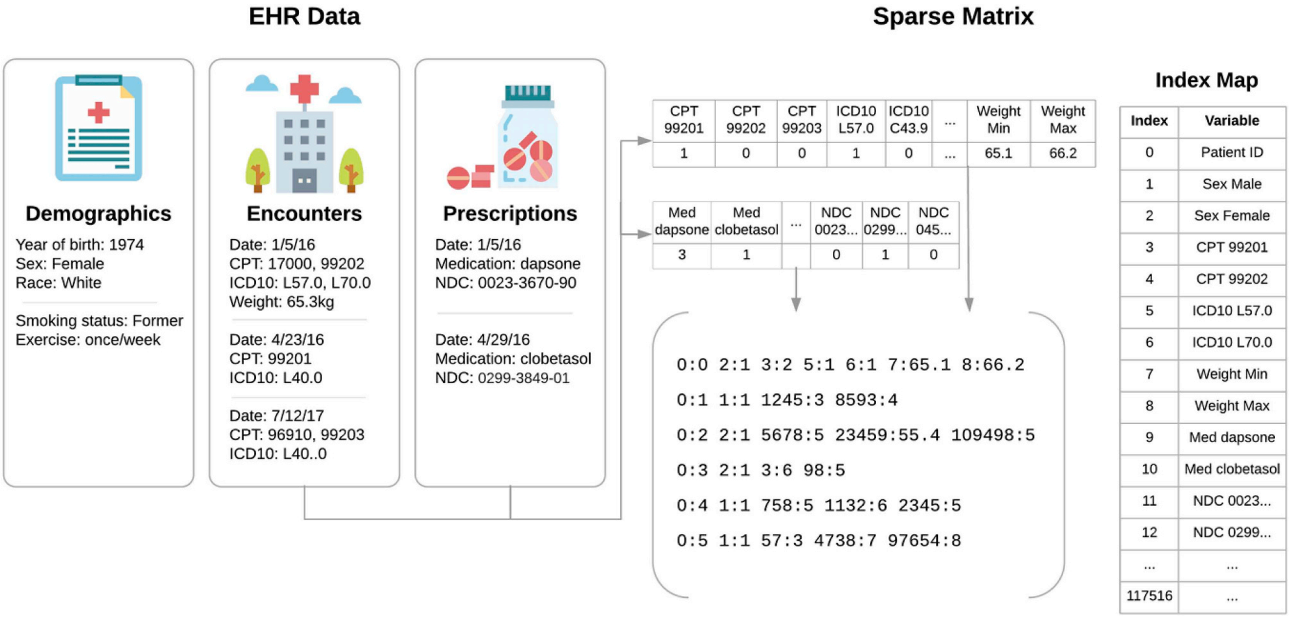


Fig. 1. Variable missingness for a random sample of patient records ordered by most present variables.



CPT 99201	CPT 99202	CPT 99203	ICD10 L57.0	ICD10 C43.9	...	Weight Min	Weight Max
1	0	0	1	0	...	65.1	66.2

Med dapsone	Med clobetasol	...	NDC 0023...	NDC 0299...	NDC 045...
3	1		0	1	0

```
0:0 2:1 3:2 5:1 6:1 7:65.1 8:66.2
0:1 1:1 1245:3 8593:4
0:2 2:1 5678:5 23459:55.4 109498:5
0:3 2:1 3:6 98:5
0:4 1:1 758:5 1132:6 2345:5
0:5 1:1 57:3 4738:7 97654:8
```

Index	Variable
0	Patient ID
1	Sex Male
2	Sex Female
3	CPT 99201
4	CPT 99202
5	ICD10 L57.0
6	ICD10 L70.0
7	Weight Min
8	Weight Max
9	Med dapsone
10	Med clobetasol
11	NDC 0023...
12	NDC 0299...
...	...
117516	...

Fig. 2. Sparse matrix creation process. First, longitudinal data are aggregated to create one vector for each patient. Then, vectors are collected into a sparse matrix using an index map to relate vector indexes to clinical variables. Note: Values in this figure are fictional and do not represent actual patients in the dataset.

aggregated for each patient by using count vectors of the data elements. For example, if a patient had four visits with a CPT code of 99201, the entry for "CPT 99201" in their vector would be "4." Because there are more than 100,000 discrete data points, most entries in the aggregated patient data were zero, resulting in a collection of sparse vectors. The count vectors also account for missing data: if a patient did not have a record of a particular feature value, the count resulted in zero. We performed mean imputation for the small number of patients that did not have a recorded year of birth. The vectors were collated and stored using a sparse matrix format, in which matrix values were represented by their index (or position) in a vocabulary, and zero values were not stored [15]. The "# Categories" column in Appendix A shows the number of new features that were added by creating count vectors of the source features, and the "Matrix Value" column corresponds to the actual number that is stored in the matrix for the patient/feature entry.

To evaluate model performance on smaller datasets and given the high class imbalance present in the training dataset, we used random undersampling (RUS) to create multiple sampled datasets [16]. RUS enabled us to keep all the positive instances (i.e., melanoma) but randomly remove negative instances (i.e., no melanoma). We evaluated models on the original dataset along with sampled datasets according to the following target positive class ratios: 0.01, 0.1, 0.25, and 0.5.

We utilized several libraries in the Python data science ecosystem to execute the experiments. We used *scikit-learn* [17] to train and evaluate

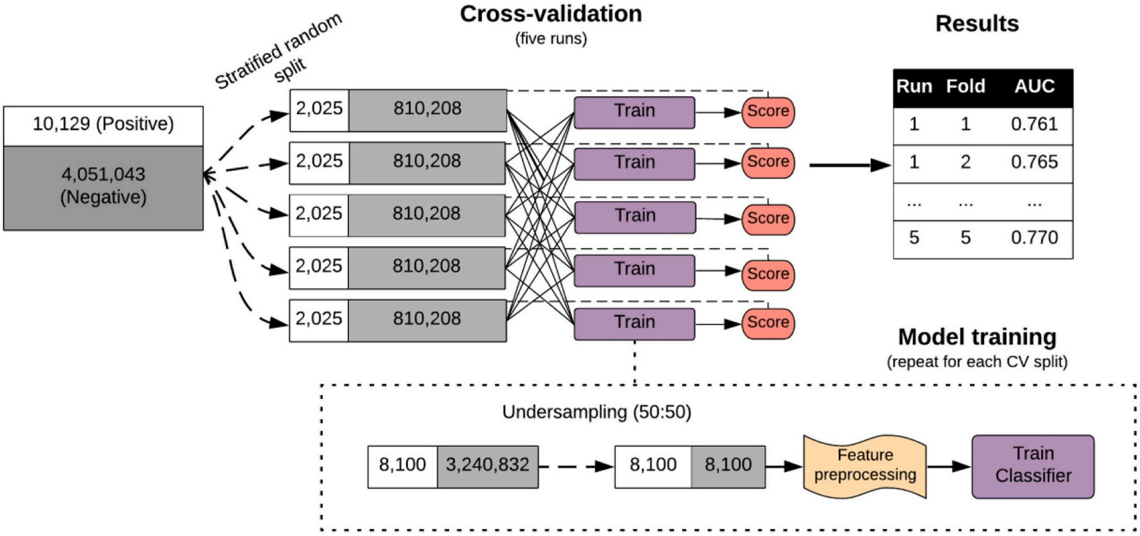


Fig. 3. Example ML Pipeline. Within each cross-validation run, the data are sampled and processed. Then, the results from each run are collected together.

all classifiers, *numpy* [18] and *scipy* [19] for matrix processing, and *imbalanced-learn* [20] to perform data sampling with *scikit-learn*.

2.3. Model training & evaluation

We built logistic regression (LR), random forest (RF), and XGBoost (XGB) models to evaluate performance across the original and sampled datasets. XGBoost is an implementation of regularized gradient-boosted trees [21]. Model hyperparameters were selected based on a grid search; LR: L1 penalty (LASSO), regularization parameter $C = 0.5$, RF: 500 trees, no maximum depth, XGB: learning rate 0.1, 500 trees, maximum depth 3. Given that LR models can be affected by high dimensionality, we selected the top 1000 features ranked according to the χ^2 statistic. We did not perform any feature selection for the RF or XGB models, because these models inherently select features.

We trained and evaluated all models using five-fold cross-validation repeated five times. An example model pipeline is provided in Fig. 3. We performed data sampling and feature preprocessing separately in each training fold rather than for the whole dataset beforehand. This resulted in 25 runs that can be used for statistical tests. We evaluated the discriminative performance of the models according to the area under the receiver operating characteristic curve (AUC). The dataset with the highest AUC for each classifier was selected for further examination with the following metrics: sensitivity (recall), specificity, precision, and area under the precision-recall curve (AUPRC). We used Welch's two-sample *t*-test to test the significance of differences in means of the results. The level of statistical significance was set at $p < 0.05$.

Because all experiments were conducted using Amazon EC2, we were able to directly calculate the cost of training each model configuration. Running time is not the best comparison across different classifiers, because the same model configuration can be run on more advanced hardware that would speed up running time. In addition, using an instance with more CPUs would only benefit models that support multithreading. Therefore, we estimated the cost of training a model as follows: [running time] × [EC2 cost for that instance]. EC2 has a pricing model that increases with hardware complexity. The instance types used along with resource specifications and cost calculation are outlined in Appendix B.

To assess the clinical utility and accuracy of the models, we examined the selected features and importance of the features for each model. The different models allow different degrees of interpretability. LR provides a global model as there is a single trained regression function with weights for each feature. These weights are used across all instances and give a sense of the importance of each feature across the entire population. Tree models such as RF and XGB provide both local and global interpretability. The tree path that an instance takes to arrive at a prediction is unique to a group of instances, not the entire population. Global interpretability can be assessed by calculating the average reduction of loss when splitting on each particular feature (Gini importance [22]). As shown by Lundberg et al., this approach can be inconsistent when comparing different models [23]. Therefore, we use Shapley Additive Explanations (SHAP) values to achieve a consistent and accurate representation of feature importance for RF and XGB, which can be used for both global and local interpretation [23,24].

3. Results

3.1. Population

There were a total of 4,061,172 patients in the MAMEL dataset that met the inclusion criteria, 10,129 of whom were diagnosed with melanoma within one year (Table 1). Compared to the “no melanoma” class, the “melanoma” class had a lower proportion of females (59.33% vs. 39.74%), and higher proportions of white race (69.57% vs. 75.26%) and family history of melanoma (11.97% vs. 13.69%).

3.2. Performance

Table 2 outlines the sizes of the original dataset and each sampled dataset as well as the average performance for the three classifiers; these results and error bars for the minimum/maximum values across all runs are plotted in Fig. 4. The greatest AUC (0.7991) was achieved by the XGB model on the 1 m dataset, but this was not significantly better than that on the original 4 m dataset (0.7988, $p = 0.846$). Training the XGB model with the 40k dataset achieved statistically comparable results to the full dataset (0.7971, $p = 0.1797$). The AUCs for the LR and XGB models were relatively unaffected by the reduction in dataset size, while the performance of the RF model actually improved when sampling was introduced. The best RF model had an AUC of 0.7736 on the 20k dataset compared to the baseline of 0.6949

Table 1
Patient population.

Variable	Value	No melanoma (*N*, %)	Melanoma (*N*, %)	*P*
Total patients		4,051,043 (99.75)	10,129 (0.25)	
Age (years)	(mean ± SD)	57.00 ± 19.88	68.31 ± 12.59	< 0.001
Sex	Female	2,403,446 (59.33)	4025 (39.74)	< 0.001
Race/ethnicity	African-American	59,367 (1.47)	–	< 0.001
	Asian	32,714 (0.81)	–	
	Hispanic	108,989 (2.69)	135 (1.33)	
	White	2,818,378 (69.57)	7623 (75.26)	
	Other	1,031,595 (25.46)	2362 (23.32)	
Geographic region	Midwest	743,581 (18.36)	1496 (14.77)	< 0.001
	Northeast	778,382 (19.21)	1621 (16)	
	South	1,700,964 (41.99)	4686 (46.26)	
	West	802,258 (19.8)	2269 (22.4)	
	Other	25,858 (0.64)	57 (0.56)	
Family history of melanoma		484,882 (11.97)	1387 (13.69)	< 0.001
History (number of visits) [a]	Mean; Median (Q1–Q3)	3.95; 2 (1–5)	5.02; 3 (1–7)	< 0.001
History (days) [b]	Mean; Median (Q1–Q3)	484.13; 371 (31–798)	557.72; 454 (40–932)	< 0.001

[a] Number of visits recorded prior to index visit.
[b] Days between earliest visit recorded and index visit.

Table 2
Average results for each dataset and classifier.

Dataset	Total (*N*)	Negative (*N*)	Positive (%)	Average AUC			Average training cost ($)		
				LR	RF	XGB	LR	RF	XGB
4 M	4,061,172	4,051,043	0.25	0.7617	0.6949	0.7988	1.1466	2.6703	0.9648
1 M	1,012,900	1,002,771	1	0.7651	0.7415	0.7991	0.1373	0.5312	0.1347
100K	101,290	91,161	10	0.7710	0.7682	0.7989	0.0152	0.0244	0.0168
40K	40,516	30,387	25	0.7713	0.7734	0.7971	0.0060	0.0111	0.0099
20K	20,258	10,129	50	0.7674	0.7736	0.7921	0.0038	0.0072	0.0068

AUC: area under the ROC curve, LR: logistic regression model, RF: random forest model, XGB: XGBoost model.

($p < 0.001$). Additional performance metrics for each best performing model are provided in Table 3. XGB had the highest AUPRC (0.0136), sensitivity (0.7529), and precision (0.0060), while LR had a slightly higher specificity (0.7045).

The RF model was the most expensive of the three classifiers, costing an average of $2.6703 for a single model fit using the full 4 m dataset. For a run of 5-fold cross-validation with 5 repeats, this would

Table 3
Additional metrics.

Classifier	Best size	AUPRC	Sensitivity (Recall)	Specificity	Precision
LR	40k	0.0100	0.6961	0.7045	0.0059
RF	20k	0.0095	0.7032	0.6952	0.0057
XGB	1 m	0.0136	0.7529	0.6877	0.0060

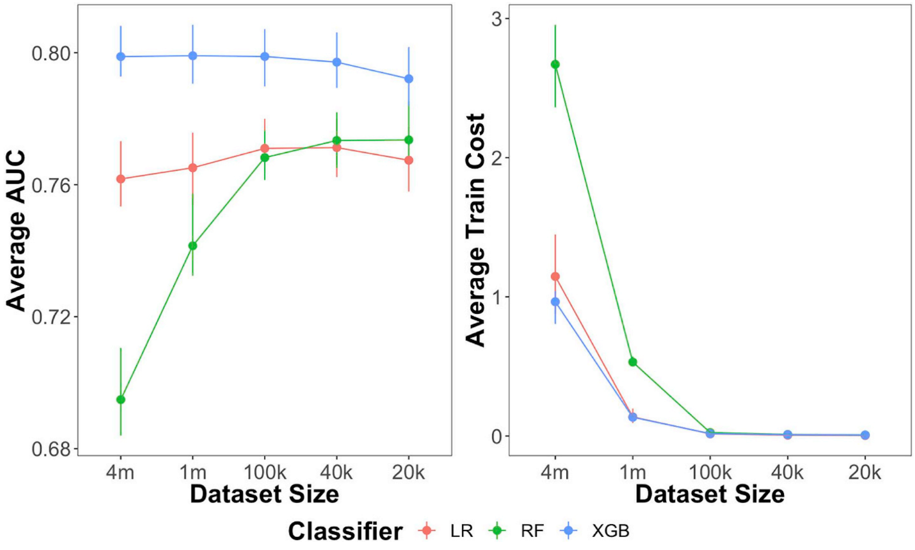


Fig. 4. Average results for each classifier and dataset.

Table 4
LR model weights (no history).

Highest weights		Lowest weights	
Intercept: 0.972721			
Feature	β	Feature	β
icd10_D48.5	3.968976	chief_complaint_Pimples (Acne)	−1.168288
icd9_173.31	1.666087	exam_chest	−0.971720
procedure_ShaveBiopsy	1.659514	icd9_V65.49	−0.930093
icd10_L57.0	1.407742	diagnosis_Acne	−0.870827
diagnosis_Basal Cell Carcinoma	1.374019	procedure_NCounselingAcne	−0.745976
procedure_mipsQuality	0.792211	static_ethnic_group_HISPANIC_OR_LATINO	−0.568793
biopsy_result_Basal Cell Carcinoma	0.642071	static_sex_FEMALE	−0.487668
procedure_PunchBiopsy	0.482086	static_race_african_american	−0.450970
icd10_Z08	0.479657	procedure_Prescription	−0.273696
cpt_11100	0.410025	diagnosis_Milia	−0.244045

cost over $66. Meanwhile, the XGB model cost $0.9648 per model fit, and the LR model cost $1.1466 per model fit on the same dataset. When run on smaller datasets, all models were significantly cheaper than the respective baselines, with datasets smaller than 1 m costing < $0.1000 per model fit. The LR model had the cheapest cost with $0.0038 for the 20k dataset, while the RF and XGB models were $0.0072 and $0.0068 for the smallest dataset, respectively.

4. Discussion

The results of the present study offer several perspectives on the intersection of risk models, EHR systems, and big data. Datasets for specific biomedical and health applications can be small because of limited data sharing between institutions, strict inclusion criteria, and a lack of structured clinical data. Risk models are often built with data collected from individual healthcare or academic institutions. While large centers can attract patients from different geographical areas, the treatment and data collection processes are still localized to the center and might not pair well with data from elsewhere. Furthermore, institutions generally do not share data, resulting in many different models being built from fragmented datasets. The dataset used in the present study is unique in that it provides over 100,000 structured data points from over 4 million dermatology patients throughout the U.S. Though the dataset is unique, the model can still have wide applicability due to the number of patients treated by physicians using the EHR system.

Distributed computing is often required to deal with big data. While packages such as Spark ML [25] provide distributed ways to train machine learning algorithms, there are many more libraries and algorithms available on non-distributed infrastructures (i.e., single machines). We found that the best approach for our scenario was to use Spark to perform data collection and transformation and then save the data into a sparse format to use with single machines. By doing so, we were able to use multiple machines to train and evaluate multiple machine learning models in parallel rather than using a cluster of machines to train one model at a time. For other datasets that are large, high-dimensional, and dense, cluster computing may still be required for model training.

Although the machine learning community often assumes that more data means better models [26], we hypothesize that this might not be true in cases with truly massive amounts of data. Here, we found that using datasets with tens of thousands of instances could achieve statistically similar (i.e., XGB models) or better (i.e., RF models) performance than when using the full dataset (n = 4,061,172). This is likely because of a high level of homogeneity among instances in the negative class, which means that less instances need to be used to produce a generalizable model. We will explore this hypothesis in future works. The RF performance increase using less data may be explained by tree overfitting in the forest. The datasets with less data had shallower trees, resulting in more generalizability when evaluating test data. Using fewer instances means that less sophisticated computing infrastructure can be used, which allows researchers to continue to use known methods and tools rather than worrying about how to handle big data in their machine learning workloads.

The most important features in the LR model are provided in Tables 4–5. We found that models trained with patients that had historical

Table 5
LR model weights (history).

Highest weights		Lowest weights	
Intercept: 1.647285			
Feature	β	Feature	β
procedure_Mohs	2.637419	procedure_SutureRemoval	−1.542473
icd10_D48.5	2.581320	follow_up_diagnosis_Acne	−1.464221
hist_icd10_L57.0	2.307007	icd9_V65.49	−1.247825
procedure_ShaveBiopsy	1.589884	static_race_african_american	−1.156283
diagnosis_Basal Cell Carcinoma	1.529410	diagnosis_Acne	−1.020633
hist_icd10_D48.5	1.517638	hist_visit_range_days	−0.925938
procedure_Defer	1.452915	icd9_706.1	−0.654565
procedure_ExcisionMalignant	1.443212	icd10_Z71.89	−0.547420
hist_icd10_Z87.2	1.376110	procedure_TreatmentRegimen	−0.497152
cpt_17004	1.255758	static_sex_FEMALE	−0.476317

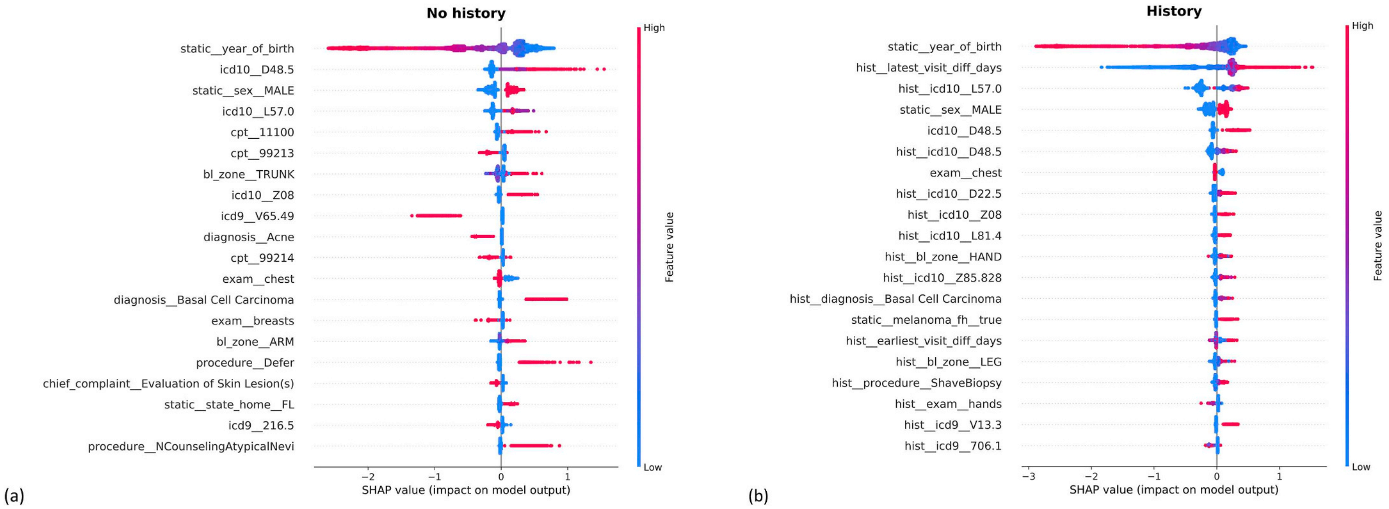


Fig. 5. XGB SHAP values from a random sample of instances for the (a) "no history" and (b) "history" populations. Each dot represents the impact of the particular feature for a given instance and is colored according to what magnitude of value contributes to the model impact. For example, a high feature value of "static_sex_MALE" has a positive impact on model output, meaning male sex is a factor that increases melanoma risk for those instances.

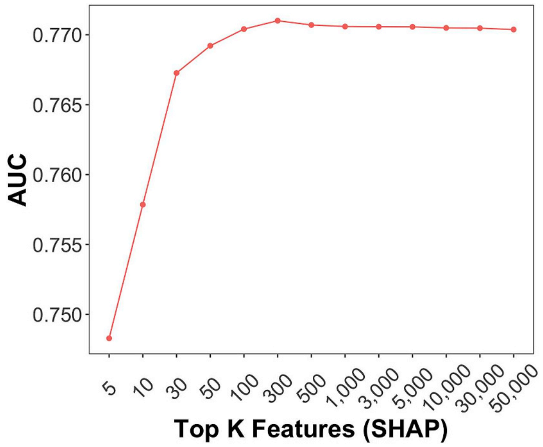


Fig. 6. Model performance as more features are included in the XGB model using the 1 m dataset. The top features are ranked according to the mean SHAP values from a trained model using all features.

visits ("history") versus those that did not ("no history") achieved similar predictive performance (AUC = 0.7597 vs. 0.7586) but utilized different features for each population. Therefore, we explore selected features separately for these populations. As we are not able to present all 1000 features due to space constraints, we display the top ten features with positive weights and top ten features with negative weights. The highest predictors for melanoma risk are the presence of other cancerous and precancerous lesions such as basal cell carcinoma or actinic keratoses, neoplasms of uncertain behavior, and history of a malignant lesion, or treatments for these conditions (Mohs surgery, excision). Negative predictors of melanoma risk are related to lower-risk populations (African American, Hispanic, Female), and dermatology diagnoses that may be an indicator of young age (acne). For patients with history, we note that the model selects several variables covering historical visits (hist_*). Age, race, and presence of other lesions are fairly known general risk factors for the cancer [27], so we can confirm that the EHR system is capturing relevant clinical information for melanoma risk prediction. The goal of this model is to be able to be used for personalized patient care; therefore, we examine more individualized feature importance using the XGB model and SHAP values.

A summary plot of the top twenty features in the XGB model according to their mean SHAP values is provided in Fig. 5. As with LR, we found that the features selected by a model on the "no history" (AUC = 0.8173) versus "history" (AUC = 0.7934) populations were different. The features selected in the XGB model have some substantial differences from the LR model. Particularly, year of birth (age) is the most important predictor, followed by actinic keratoses (L57.0), neoplasm of uncertain behavior (D48.5), and sex. Other risk factors not present in the LR model are melanoma family history, evaluations of various body locations (trunk, leg, chest, hands) and geographic location (home address in Florida). The plot provides an estimate of both individualized and global feature importance by plotting the SHAP values for a random sample of instances. Older age (lower year of birth) has the largest positive impact on risk for both populations, as well as different magnitudes of negative risk decreasing as age decreases. The time between the index visit and the patient's most recent visit ("hist_latest_visit_diff_days") was the second most important factor for the "history" population, and seems to have an increasing impact on risk as the time increases. The top features for the "history" model are almost exclusively from historical visit data rather than the index visit. This shows that history is indeed important for building estimating risk, but other factors can also be used from the index visit if history is not available. These observations show that a simple global model does not necessarily provide the best estimate of melanoma risk.

Age has an impact across the largest group of patients, but the other features appear to have effects only for localized groups, meaning that a large number of features must be included to produce the most effective model. To evaluate this hypothesis, we trained an XGB model on the 1 m dataset with an increasing number of the most important features according to their SHAP values (Fig. 6). We see that 300 features are required to achieve the best performing model, meaning that hundreds of different factors from the patient's history can affect their risk of the disease. Deployment into the structured EHR system is ideal for this type of model, as a patient's risk can be evaluated in real-time rather than using an external risk evaluation tool to manually input data.

Limitations of this dataset include selection bias and data quality. Because all patients in this study have already visited a dermatologist, we might be missing key patients who do not regularly visit a dermatologist. Increased interoperability and data sharing between institutions can help reduce this limitation, which is a key goal of the 21st Century Cures Act [9]. There is, however, clinical utility of this model on a dermatology population. We evaluated the chief complaint of the earliest visit for each patient in the dataset and found that 1,090,042 (26.91%) patients in the "no melanoma" class had a chief complaint for other conditions, such as acne, verruca vulgaris, or a rash. Even in the "melanoma" class, 1167 (11.52%) of patients did not initially present for a skin check. These patients are ideal candidates for our model, as they may be high-risk for melanoma and not know it. While some EHR systems provide a structured data input solution, variable input might differ substantially among providers, limiting the depth of data available. Accordingly, the size of the current dataset helps to alleviate concerns regarding consistency and missing data. While many factors selected by these models indicate patients that may be already presenting for skin checks, the model can still provide value as the predictions are personalized for each patient. Furthermore, we only explored the top 20 features for each model, but over 300 are required to produce the best model. The most informative features for specific predictions may involve other features not routinely associated with melanoma risk.

5. Conclusion

We described a case study of learning from big data to produce an effective melanoma risk prediction model based on data collected from a large representative dermatology EHR system covering millions of patients across the U.S. Our study provides a reference framework for machine learning studies using large, high-dimensional, and imbalanced EHR data. We used a distributed processing infrastructure for collecting and formatting the data as well as a non-distributed infrastructure for machine learning. Then, we achieved statistically similar or better performance using a sampled dataset versus the original data, saving hundreds of dollars in cloud computing costs for model experimentation.

The structured-data EHR and cloud-based model training process described herein addressed the shortcomings identified in previous cancer risk modeling studies. *Availability of structured clinical data*: the structured, cloud-based EHR system provided consistently collected data points across millions of patients at different practices. *Old data*: the data consistency allowed for rapid querying, de-identification and transformation of data to use for training machine learning models. The time between the end of the study period (December 2017) and study completion was about one year. *Advanced modeling methods*: we used a familiar and feature-rich machine learning framework (*scikit-learn)* with advanced machine learning techniques such as random forest, XGBoost, data sampling, and feature selection.

Future studies should aim to validate the data infrastructure choices on other clinical datasets and improve the accuracy of the melanoma risk models. More advanced algorithms such as artificial neural networks can be explored to take advantage of the longitudinal data available in EHR systems. While this current study does not utilize image data, future research should consider combining both EHR and image data to provide the best risk models for dermatology patients.

Author contributions

ANR and TMK created the experimental designs and facilitated data acquisition. ANR performed the data processing, experiments, and drafted the initial manuscript. TMK provided manuscript edits and guidance for analysis and interpretation of the results.

Conflict of interest statement

None declared.

Acknowledgements

We would like to thank the anonymous reviewers and editor for their insightful feedback and helpful comments. Also, we thank the reviewers at the Data Mining and Machine Learning Laboratory at Florida Atlantic University. The icons in Fig. 2 were made by *smalllikeart* on www.flaticon.com.

Appendix A. Data elements

Data Elements.

Group	Name	Description	% Missing	# Categories	Matrix value
Patient data	static_year_of_birth	Year the patient was born	< 0.1%	–	(integer)
	static_sex	Birth sex (male, female, other)	0.00%	3	One-hot encoded categorical
	static_ethnic_group	Ethnic group (Hispanic or Latino, not Hispanic or Latino, other)	0.00%	3	One-hot encoded categorical
	race	Race (African-American, Asian, White, other)	21.20%	4	One-hot encoded categorical
	static_state_home	U.S. state of home address (including D.C.)	0.50%	51	One-hot encoded categorical
	static_melanoma_fh	If patient has a family history of melanoma	3.20%	–	(0/1)
	static_allergy	Drug allergen descriptions [28]	67.00%	7406	(integer) Number of times allergy was recorded across all visits
	static_family_history_snomed	Family history SNOMED codes	70.60%	418	(integer) Number of times family history condition was recorded across all visits
Visit data	cpt	Standard codes describing medical procedures for billing purposes [29]	7.80%	577	(integer) Number of units billed for each CPT code
	icd9	Diagnostic codes used for disease classification and billing (9th edition [30])	7.90%	896	(integer) Number of times each ICD9 code was referenced in the bill
	icd10	Diagnostic codes used for disease classification and billing (10th edition [30])	7.90%	2819	(integer) Number of times each ICD10 code was referenced in the bill
	snomed	Standardized medical terminology covering terms beyond only procedures or diagnoses [31]	38.90%	180	(0/1) If SNOMED code was associated with the visit
	loinc	Identifiers for laboratory orders [32]	98.20%	1953	(integer) Number of times each LOINC code was ordered in the visit
	cash_charge	Direct charges to the patient for non-medical procedures (categories represent a diagnosis/procedure combination that was charged for)	96.00%	5477	(float) Dollar amount of cash charges in the bill
	vital	Height, weight, temperature, blood pressure (systolic/diastolic), pulse, respiration	90.80%	7	(float) Numeric value of each measurement
	ros	Series of questions to identify symptoms the patient is presenting with (ex. problems with healing, rash, hay fever, sore throat)	79.40%	131	(0/1) If ROS question response was "yes"
	chief_complaint	Reason why the patient is visiting the dermatologist (ex. skin lesion, skin lesion follow up, rash, acne)	29.10%	405	(0/1) If each chief complaint was documented in the visit
	follow_up_diagnosis	Previous diagnosis the patient is following up on	64.90%	1700	(0/1) If each diagnosis was documented as the follow-up diagnosis for the visit
	exam	Body elements that the physician examined (ex. scalp, head, chest, neck, back)	9.10%	95	(0/1) If each body element was examined
	diagnosis	Findings noted in the exam to associate procedures with (ex. benign nevi, psoriasis, acne, melanoma, history of melanoma)	0.20%	2048	(integer) Number of times each diagnosis was documented in the visit
	procedure	Procedures and plans performed during the visit (ex. liquid nitrogen, counseling, reassurance, biopsy)	0.50%	2117	(integer) Number of times each procedure was documented in the visit
	bl_zone	Body locations associated with a finding and/or procedure (ex. head, face, trunk, scalp, leg)	6.40%	64	(integer) Number of times each body location was documented in the visit
	biopsy_result	Result of a biopsy/excision performed in the visit (ex. dysplastic nevus, basal cell carcinoma, melanoma)	87.40%	708	(integer) Number of times each result was received
	medication	Medication name of a prescription written during the visit	72.00%	1107	(integer) Number of times each medication was prescribed during the visit
	ndc	NDC code of a prescription written during the visit [33]	72.00%	5086	(integer) Number of times each NDC was prescribed during the visit

Historical visit data	hist_num_visits	Number of visits documented before the current visit	0.00%	–	(integer)
	hist_earliest_visit_diff_days	Number of days between earliest historical documented visit and current visit	0.00%	–	(integer)
	hist_latest_visit_diff_days	Number of days between latest historical visit documented and current visit	0.00%	–	(integer)
	hist_visit_range_days	Number of days between earliest and latest historical visit	0.00%	–	(integer)
	hist_cpt	Standard codes describing medical procedures for billing purposes [29]	23.20%	905	(integer) Number of times each CPT code was referenced across all historical visits
	hist_icd9	Diagnostic codes used for disease classification and billing (9th edition [30])	23.60%	1445	(integer) Number of times each ICD9 code was referenced across all historical visits
	hist_icd10	Diagnostic codes used for disease classification and billing (10th edition [30])	26.80%	3212	(integer) Number of times each ICD10 code was referenced across all historical visits
	hist_snomed	Standardized medical terminology covering terms beyond only procedures or diagnoses [31]	44.10%	244	(integer) Number of times each SNOMED code was referenced across all historical visits
	hist_loinc	Identifiers for laboratory orders [32]	94.70%	3899	(integer) Number of times each LOINC code was ordered across all historical visits
	hist_cash_charge	Direct charges to the patient for non-medical procedures (categories represent a diagnosis/procedure combination that was charged for)	93.30%	6602	(float) Dollar amount of cash charges in the bill across all historical visits
	hist_vital	Height, weight, temperature, blood pressure (systolic/diastolic), pulse, respiration	87.30%	49	(float) Min/max/mean/median/std of the numeric values of each measurement across all historical visits
	hist_ros	Series of questions to identify symptoms the patient is presenting with	72.60%	129	(integer) Number of times each ROS question response was “yes” across all historical visits
	hist_chief_complaint	Reason why the patient is visiting the dermatologist	22.30%	425	(integer) Number of times each chief complaint was documented across all historical visits
	hist_follow_up_diagnosis	Previous diagnosis the patient is following up on	56.10%	1826	(integer) Number of times each diagnosis was documented as the follow-up diagnosis across all historical visits
	hist_exam	Body elements that the physician examined	22.90%	104	(integer) Number of times each body element was examined across all historical visits
	hist_diagnosis	Findings noted in the exam to associate procedures with	21.60%	2143	(integer) Number of times each diagnosis was documented across all historical visits
	hist_procedure	Procedures and plans performed during the visit	21.60%	2206	(integer) Number of times each procedure was documented across all historical visits
	hist_bl_zone	Body locations associated with a finding and/or procedure	23.30%	75	(integer) Number of times each body location was documented across all historical visits
	hist_biopsy_result	Result of a biopsy/excision performed in the visit	68.60%	797	(integer) Number of times each result was received across all historical visits
	hist_medication	Medication name of a prescription written during the visit	52.80%	1833	(integer) Number of times each medication was prescribed across all history
	hist_ndc	NDC code of a prescription written during the visit [33]	54.30%	9224	(integer) Number of times each NDC was prescribed across all history

Appendix B. EC2 instance types and cost calculation

EC2 Instance Types

Instance type	CPUs	Memory (GB)	**Hourly** Price ($)	Models
r4.2xlarge	8	61	0.532	LR
c4.8xlarge	36	60	1.591	RF
				XGB

LR: logistic regression, RF: random forest, XGB: XGBoost.

Instance and cost data as of June 27, 2018. Amazon offers discounted rates by using Spot instances, but those prices are not constant over time, so we used the on-demand hourly rate for comparisons. We calculated the train cost for each model fit according to the following formula:

$$Train\ cost = \frac{Train\ time}{3600} * Hourly\ price$$

Where *train time* is the average time (in seconds) spent training a single model (i.e., one training fold in cross validation) and *hourly price* is the Amazon EC2 hourly cost.

References

[1] National Cancer Institute, Cancer Statistics, (2018) https://www.cancer.gov/about-cancer/understanding/statistics , Accessed date: 15 August 2018.

[2] E.W. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, first ed., Springer-Verlag New York, 2009.

[3] A.N. Richter, T.M. Khoshgoftaar, A review of statistical and machine learning methods for modeling cancer risk using structured clinical data, Artif. Intell. Med. 90 (2018) 1–14, https://doi.org/10.1016/j.artmed.2018.06.002.

[4] J.A. Usher-Smith, J. Emery, A.P. Kassianos, et al., Risk prediction models for melanoma: a systematic review, Canc. Epidemiol. Biomark. Prevent. 23 (2014) 1450–1463, https://doi.org/10.1158/1055-9965.EPI-14-0295.

[5] A.C. Society, Cancer Facts & Figures 2019, American Cancer Society, Atlanta, GA, 2019.

[6] A. Romero-Lopez, X. Giro-i-Nieto, J. Burdick, et al., Skin lesion classification from dermoscopic images using deep learning techniques, Biomedical Engineering, ACTAPRESS, Innsbruck, Austria, 2017, , https://doi.org/10.2316/P.2017.852-053.

[7] A. Esteva, B. Kuprel, R.A. Novoa, et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118, https://doi.org/10.1038/nature21056.

[8] J. AK, Meaningful use of electronic health records: the road ahead, JAMA 304 (2010) 1709–1710, https://doi.org/10.1001/jama.2010.1497.

[9] K.L. Hudson, F.S. Collins, The 21st century cures act — a view from the NIH, N. Engl. J. Med. 376 (2017) 111–113, https://doi.org/10.1056/NEJMp1615745.

[10] S. Doan, M. Conway, T.M. Phuong, et al., Natural language processing in biomedicine: a unified system architecture overview, Clin. Bioinformat. (2014) 275–294.

[11] A.N. Richter, T.M. Khoshgoftaar, Modernizing Analytics for melanoma with a large-scale research dataset, IEEE International Conference on Information Reuse and Integration (IRI), IEEE, San Diego, CA, 2017, , https://doi.org/10.1109/IRI.2017.45 2017. 551–8.

[12] Methods for De-identification of PHI | HHS.gov, HHS.gov. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/ , Accessed date: 18 March 2017.

[13] A. Avati, K. Jung, S. Harman, et al., Improving palliative care with deep learning, 171106402, (2017).

[14] M. Zaharia, R.S. Xin, P. Wendell, et al., Apache spark: a unified engine for big data processing, Commun. ACM 59 (2016) 56–65.

[15] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Transact. Intell. Sys. Technol. 2 (2011) 27 1–27:27.

[16] J. Van Hulse, T.M. Khoshgoftaar, A. Napolitano, Experimental perspectives on learning from imbalanced data, Proceedings of the 24th International Conference on Machine Learning. ACM, 2007, pp. 935–942 http://dl.acm.org/citation.cfm?id=1273614 , Accessed date: 31 May 2016.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[18] S. van der Walt, S.C. Colbert, G. Varoquaux, The NumPy array: a structure for efficient numerical computation, Comput. Sci. Eng. 13 (2011) 22–30, https://doi.org/10.1109/MCSE.2011.37.

[19] E. Jones, T. Oliphant, P. Peterson, SciPy: Open Source Scientific Tools for Python, (2014).

[20] G. Lemaître, F. Nogueira, C.K. Aridas, Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (2017) 1–5.

[21] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: : ACM, 2016, pp. 785–794, , https://doi.org/10.1145/2939672.2939785.

[22] B.H. Menze, B.M. Kelm, R. Masuch, et al., A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data, BMC Bioinf. 10 (2009) 213, https://doi.org/10.1186/1471-2105-10-213.

[23] S.M. Lundberg, G.G. Erion, S.-I. Lee, Consistent individualized feature attribution for tree ensembles, arXiv:180203888 [cs, stat] (11 February 2018) , Accessed date: 27 February 2019http://arxiv.org/abs/1802.03888.

[24] Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. ;:10.

[25] X. Meng, J. Bradley, B. Yavuz, et al., MLlib: machine learning in Apache spark, J. Mach. Learn. Res. 17 (2016) 1235–1241.

[26] T. van der Ploeg, P.C. Austin, E.W. Steyerberg, Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints, BMC Med. Res. Methodol. 14 (2014) 137.

[27] C.G. Watts, M. Dieng, R.L. Morton, et al., Clinical practice guidelines for identification, screening and follow-up of individuals at high risk of primary cutaneous melanoma: a systematic review, Br. J. Dermatol. 172 (2015) 33–47, https://doi.org/10.1111/bjd.13403.

[28] S. Liu, W. Ma, R. Moore, et al., RxNorm: prescription for electronic drug information exchange, IT Professional 7 (2005) 17–23.

[29] AAPC, What is HCPCS? https://www.aapc.com/resources/medical-coding/hcpcs.aspx , Accessed date: 9 April 2017.

[30] Organization WH, others, International Classification of Diseases (ICD), (2012).

[31] SNOMED International, SNOMED CT: The global language of healthcare, http://www.snomed.org/snomed-ct , Accessed date: 9 April 2017.

[32] LOINC. LOINC: The freely available standard for identifying health measurements, observations, and documents. https://loinc.org/(accessed 9 Apr 2017).

[33] U.S. Food & Drug Administration, National drug code directory, https://www.fda.gov/drugs/informationondrugs/ucm142438.htm , Accessed date: 9 April 2017.