Analysis of the Elderly's Internet Accessed Time using XGB Machine Learning Model for Solving the Level of the Information Gap of the Elderly

Hung Viet Nguyen[1]
Department of Artificial Intelligence
College of Computer Engineering, Inje University
Gimhae 50834, Republic of Korea

Haewon Byeon[2]*
Department of Digital Anti-aging Healthcare
Graduate School (BK-21), Inje University
Gimhae 50834, Republic of Korea

***Abstract*—This study aims to construct machine learning models to predict the elderly's internet-accessed time. These models can resolve the information gaps in the present and future by analyzing information use factors such as internet access and mobile device usability. We analyzed 2,300 adults 55 years of age and older who participated in the national survey. This study followed a pipeline of five steps: primary data selection, data imputation to process missing data, feature ranking to identify most important features, machine learning algorithms to develop classifier models, and model evaluation. We applied the Extremely Randomized Trees classifier (Extra Tree) model, the Random Forest classifier (RF) model, and the Extreme Gradient Boosting classifier (XGB) model to look for feature ranking, then select feature importance. All classification models used the accuracy score to calculate the effect. In our study, the most accurate model for predicting the Internet access time of the elderly was the XGB model. The evaluation scores of the XGB machine learning model are very positive and bring high expectations. To solve the information gap of the elderly problem, we can use these effective models to predict the elderly object. Then, we can give some solutions to help them in a society with a strong information technology base.**

Keywords*—*Information gap; machine learning; prediction model; elderly

I. Introduction

Articles related to the fourth industrial revolution, artificial intelligence (AI), robotics, autonomous vehicles, and unscrewed aerial vehicles appear in the media daily. As such, modern society is an information society. How does an information society affect the daily life of the elderly? Research began with these questions. Information Society is the development of information and communication technology.

It refers to a society in which valuable information can be created. It means that the center of existing economic activity is shifted from goods to information, services, and knowledge. Information becomes a vital resource as much as material or energy resources. In other words, through collecting, producing, processing, and storing information, the distribution of information is spread, and this is a society in which these actions are universal. With the rapid development of information and communication technology day by day, it is rapidly entering into human life and making human life more convenient.

However, in a developing society, there are two people classes, one that does not have a computer or mobile device to access the internet, and one has those devices but cannot use it or has a low level of usage. They are the information-vulnerable class or the information-poor class [1, 2], and the representative targets are the disabled, the elderly, the low-income class, and farmers and fishers [3, 4]. An information gap is created between those who have access to new forms of information technology and those who do not. This information gap is expanding from the quantitative aspect of simply owning the internet or mobile device to the view of inequality among members of society arising from the qualitative aspect related to information literacy ability. Factors that cause this information gap include economic factors that can possess information devices, sociodemographic factors such as gender, age, race, and region, and cultural factors such as information literacy ability [5, 6, 7, 8, 9].

In this study, the level of the information gap of the elderly is predicted by the recent internet accessed time, and the factor recent internet accessed time is analyzed in terms of mobile usability factors. This study aims to construct machine learning models to predict the elderly's internet accessed time. These models can resolve the information gap in the present and future by analyzing information use factors such as internet access and mobile devices usability.

II. Methods and Materials

A. Research Subjects

The data source for this study was the 2019 Digital Information Gap Survey. The number of respondents who participated in 2019 Digital Information Gap Survey is 15,000 people aged seven and over nationwide. A detailed description of the data source is presented in Choi (2020) [10]. We analyzed 2,300 adults aged 55 or older among the subjects who completed the survey.

B. Research Process

The programming language used in this research was Python version 3.7. This study followed a pipeline of five steps: primary data selection, data imputation to process

*Corresponding Author.

missing data, feature ranking to identify most important features, machine learning algorithms to develop classifier models, and model evaluation. The primary dataset had 2300 samples, with many missing values in 233 features. This study selected thirteen features involved in mobile devices usabilities of the elderly, which are encode by "code3 (whether or not you have a smartphone; 1=yes, 2=no), code7 (internet availability; 1=yes, 2=no), code15 (availability of mobile devices (e.g. display/sound/security/alarm); 1=not at all, 4=most are available), code16 (availability of mobile devices (e.g. wifi); 1=not at all, 4=most are available), code17 (whether you can move files from mobile device to computer; 1=not at all, 4=most are available), code18 (whether you can send files/photos from mobile device to others; 1=not at all, 4=most are available), code19 (whether necessary apps can be installed/deleted/updated on mobile devices; 1=not at all, 4=most are available), code20 (whether it can scan/repair the mobile device's malicious code (virus, spyware, etc.); 1=not at all, 4=most are available), code21 (whether you can write documents or materials (memo, word, etc.) on mobile device; 1=not at all, 4=most are available), code22 (whether you can connect and communicate with others over the Internet; 1=not at all, 4=most are available), code23 (whether you can actively exchange opinions on political and social issues or problems using the Internet; 1=not at all, 4=most are available), code24 (Whether you can protect yourself from personal information exposure; 1=not at all, 4=most are available), code25 (Whether you can be responsible for the use of the Internet; 1=not at all, 4=most are available)". The target variable was defined as recent internet use experience (1=within the last month, 2=over a month, 3=never used). The dataset table is presented in Fig. 1.

	code3	code7	code15	code16	code17	code18	code19	code20	code21	code22	code23	code24	code25	code26
0	1	1	3	3	3	3	3	3	3	3	3	3	3	1
1	1	1	3	2	2	3	3	2	2	3	2	2	3	1
2	1	1	2	2	3	3	2	2	2	3	2	2	2	1
3	1	1	2	2	2	3	2	2	3	1	1	1	1	1
4	1	1	4	4	4	4	4	3	3	3	3	3	3	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2295	3	2	2	2	2	1	1	2	2	1	2	2	1	3
2296	3	2	1	2	1	1	1	2	2	2	1	2	2	3
2297	3	2	3	2	2	2	2	1	3	2	1	1	2	2
2298	3	2	3	3	3	3	4	3	3	2	1	2	2	2
2299	3	2	2	2	1	1	1	1	1	1	2	2	2	3

2300 rows × 14 columns

Fig. 1. Dataset Table.

code3 (whether or not you have a smartphone; 1=yes, 2=no), code7 (internet availability; 1=yes, 2=no), code15 (availability of mobile devices (e.g. display/sound/security/alarm); 1=not at all, 4=most are available), code16 (availability of mobile devices (e.g. wifi); 1=not at all, 4=most are available), code17 (whether you can move files from mobile device to computer; 1=not at all, 4=most are available), code18 (whether you can send files/photos from mobile device to others; 1=not at all, 4=most are available), code19 (whether necessary apps can be installed/deleted/updated on mobile devices; 1=not at all, 4=most are available), code20 (whether it can scan/repair the mobile device's malicious code (virus, spyware, etc.); 1=not at all, 4=most are available), code21 (whether you can write documents or materials (memo, word, etc.) on mobile device; 1=not at all, 4=most are available), code22 (whether you can connect and communicate with others over the Internet; 1=not at all, 4=most are available), code23 (whether you can actively exchange opinions on political and social issues or problems using the Internet; 1=not at all, 4=most are available), code24 (Whether you can protect yourself from personal information exposure; 1=not at all, 4=most are available), code25 (Whether you can be responsible for the use of the Internet; 1=not at all, 4=most are available), code26 (recent internet use experience; 1=within the last month, 2=over a month, 3=never used).

This study applied the Extremely Randomized Trees classifier (Extra Tree) model, Random Forest classifier (RF) model, and Extreme Gradient Boosting classifier (XGB) model to look for feature ranking then select feature importance [11, 12, 13]. Extreme gradient boosting (XGB) - a supervised Machine Learning (ML) algorithm was compared to Gradient Boosting classifier (GBM) model, K-Nearest Neighbors classifier (KNN) model, Random Forest (RF) model, and Extra Tree model then applied to make the most effective model with tuned hyperparameters. Three different feature ranking strategies were used for each model to determine the best combination of feature ranking techniques, number of features, and prediction model. A block diagram of the working process is shown in Fig. 2.

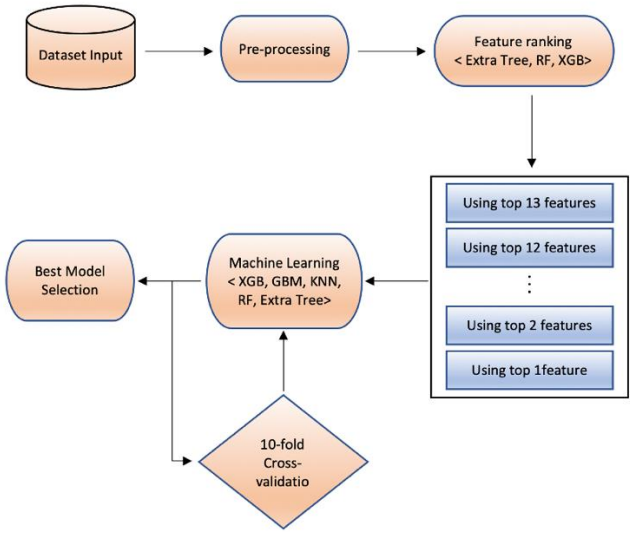


Fig. 2. Block Diagram of the Working Process.

C. Model Evaluation

All classification models used the accuracy score to calculate the effect. Models' accuracy can be defined as the relationship between true positives and true negatives.

Besides the accuracy score, this study also used Area Under the Curve (AUC) score to evaluate the model in case of which model's accuracy score is equal to the others' accuracy score. AUC measures the area beneath the ROC curve and is scale-invariant. It is also threshold invariant. AUC measures how good a model is at predicting True Positives and False Positives [14]. AUC ranges in value from 0 to 1. One model which mis-predicts 100% has an AUC of 0.0; one which predicts 100% correctly has an AUC of 1.0.

All machine learning models for prediction, augmentation, and feature ranking were developed using Python 3.7 codes that utilized the Scikit-Learn machine learning library. The best model's hyperparameters were tuned using the HyperOpt library and the stratified k-fold cross-validation, where the value of k was 10.

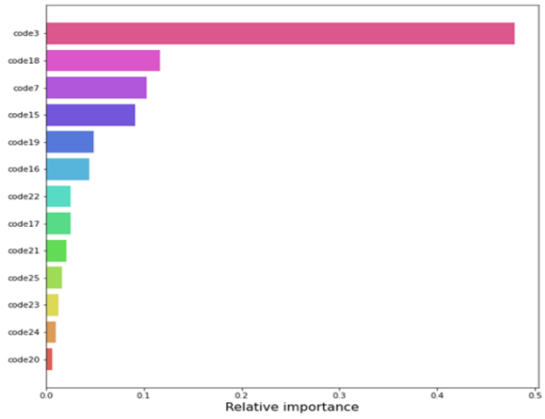
III. Results and Discussion

A. Performance Evaluation of Machine Learning Models

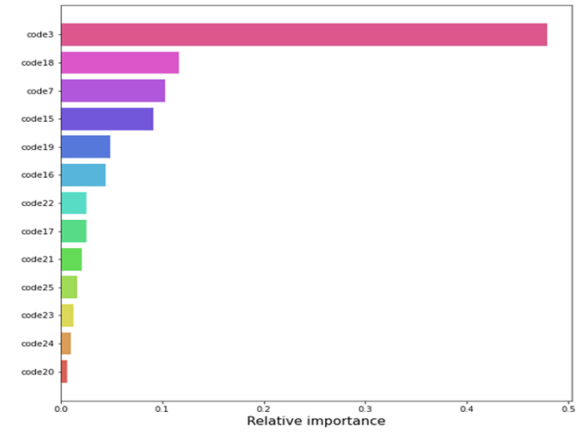
Features' correlations were observed before finding feature ranking. Fig. 3 represents the correlation heatmap of the dataset. We can see that target data, code 26, has positive relations to code 3, code 7, and has negative relations to code 15, code 16, code 18, and code 19.

Fig. 3. Features Correlation Heatmap of the Dataset.

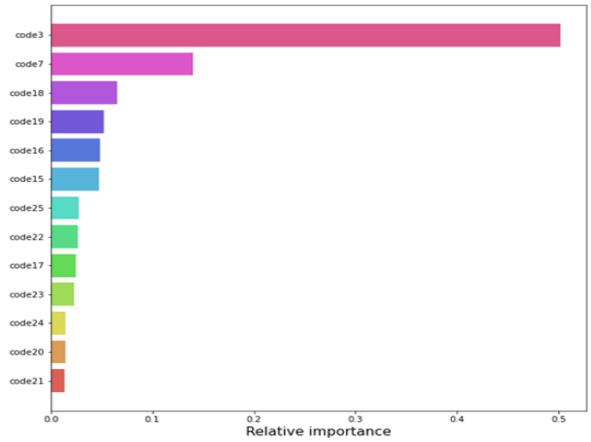
In this study, three different methods of feature ranking were applied (XGB, RF, and Extra Tree) [13]. As shown in Fig. 4, code3, code7, and code18 were the most important features in most cases. All three algorithms returned code3 as the most important feature. Code 3 represents the question about usable phone type at home, and code 7 represents the question about availability for using the internet at home. Code 18 was encoded to ask about users' ability to send files/photos from their mobile device to others.



(a)



(b)



(c)

Fig. 4. Feature Ranking (a) Extra Tree Algorithm; (b) XGB Algorithm; (c) RF Algorithm.

code3 (whether or not you have a smartphone; 1=yes, 2=no), code7 (internet availability; 1=yes, 2=no), code15 (availability of mobile devices (e.g. display/sound/security/alarm); 1=not at all, 4=most are available), code16 (availability of mobile devices (e.g. wifi); 1=not at all, 4=most are available), code17 (whether you can move files from mobile device to computer; 1=not at all, 4=most are available), code18 (whether you can send files/photos from mobile device to others; 1=not at all, 4=most are available), code19 (whether necessary apps can be installed/deleted/updated on mobile devices; 1=not at all, 4=most are available), code20 (whether it can scan/repair the mobile device's malicious code (virus, spyware, etc.); 1=not at all, 4=most are available), code21 (whether you can write documents or materials (memo, word, etc.) on mobile device; 1=not at all, 4=most are available), code22 (whether you can connect and communicate with others over the Internet; 1=not at all, 4=most are available), code23 (whether you can actively exchange opinions on political and social issues or problems using the Internet; 1=not at all, 4=most are available), code24 (Whether you can protect yourself from personal information exposure; 1=not at all, 4=most are available), code25 (Whether you can be responsible for the use of the Internet; 1=not at all, 4=most are available), code26 (recent internet use experience; 1=within the last month, 2=over a month, 3=never used.

KNN and several Tree based ML models (RF algorithm, extra tree classifier, GBM, and XGBoost) were applied to analyze our dataset using the top feature, then the top two features, the top three features, etc., continuing for all 13 features [15]. This step identified the best combination of feature ranking model and a minimum number of features to achieve the best performance. As a fundamental machine learning algorithm, K Nearest Neighbor is a method of predicting output values based on a set of input values. It is one of the least complex machine learning algorithms. It classifies the data point on how its neighbor is classified. The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. The Random Forest is used in this study because it is stable and easy to implement and provides several interesting properties, including computing variables with excellent efficiency [16]. The extra tree classifier combines the results of multiple de-correlated decision trees collected into a "forest" to produce a classification result. It has been applied in this study because it is similar to RF; however, its construction method in the forest is optimal and faster than RF [13, 16, 17]. GBM and XGBoost are two popular techniques for ensemble ML, and their performance is good on structured and tabular data. These techniques are used to solve real-life data science problems and solve them with parallel tree boosting. While both XGB and GBM use gradient boosting as a principle, there are differences in the modeling details. Specifically, XGBoost uses a more formalized formalization to control overfitting, making it perform better [18]. These techniques were used because their impact has been widely recognized in many machine learning and data mining challenges [13].

Table I shows the performance of the top models for each of the five ML algorithms that employ three feature ranking approaches utilizing three feature ranking approaches. The XGB model using the XGB feature ranking method produced the best results. This model used 11 features, reaching 0.970 of accuracy score and 0.9894 of AUC score. For this model, selected variables included code3, code15, code18, code7, code25, code22, code23, code19, code21, code16, code20, code24, and code17. GBM came in second with an accuracy of 0.970, and its AUC score of 0.9884 is slightly lower than that of XGB. Therefore, the XGB model was chosen as the best model, and its hyper-parameters were tuned to get the most effective model.

B. Performance Evaluation of XGB Classifier Model

The best model obtained in this study is the XGB model with the XGB feature ranking method. The XGB Classifier model used 11 important selected features (code 3, code 7, code 15, code 18, code 19) as feature variables and code26 as the target variable. Feature and target variables were split to 70% for training and 30% for test. The XGB model's hyperparameters were tuned by the HyperOpt library. Developed by James Bergstra, Hyperopt is a powerful Python library for hyperparameter optimization. Hyperopt uses a form of Bayesian optimization to find the best parameters for a given model. It can optimize a model with hundreds of parameters [19]. After tuning the hyper-parameters, the best XGB Classifier's hyper-parameters are 'colsample_bytree': 0.66, 'gamma': 1.06, 'learning_rate': 0.43, 'max_depth': 6, 'min_child_weight': 1.0, 'n_estimators': 14, 'subsample': 0.83. The most effective XGB model had a 0.97 accuracy score and 0997 AUC score. The model can be evaluated to work extremely effectively.

TABLE I. ANALYZING THE PERFORMANCE OF DIFFERENT MACHINE LEARNING MODELS

Algorithm	Feature Selection Models	Number of features	Accuracy	Average Recall	Average Precision	Average F1 score	Average ROC_AUC
XGB	Extra Tree	12	0.968	0.968	0.968	0.967	0.989
	Random Forest	13	0.968	0.968	0.968	0.967	0.989
	XGB	11	0.970	0.970	0.970	0.969	0.989
GBM	Extra Tree	12	0.968	0.968	0.969	0.967	0.990
	Random Forest	10	0.968	0.968	0.968	0.967	0.988
	XGB	11	0.970	0.970	0.970	0.969	0.989
KNeighbors	Extra Tree	5	0.959	0.959	0.959	0.958	0.972
	Random Forest	2	0.962	0.962	0.962	0.962	0.949
	XGB	3	0.964	0.964	0.963	0.963	0.975
Random Forest	Extra Tree	11	0.964	0.964	0.963	0.963	0.986
	Random Forest	9	0.967	0.967	0.967	0.966	0.985
	XGB	4	0.965	0.965	0.965	0.964	0.989
ExtraTree	Extra Tree	3	0.962	0.962	0.962	0.961	0.982
	Random Forest	4	0.965	0.965	0.965	0.964	0.985
	XGB	4	0.965	0.965	0.965	0.964	0.985

IV. Conclusion

In this study, the level of the information gap of the elderly can be predicted as the recent internet accessed time, and the factor recent internet accessed time is analyzed in terms of mobile usability factors. This study constructed machine learning models to predict the elderly's internet accessed time through the elderly's mobile usability factors. XGB algorithm was used to design the machine learning model (XGB Classifier Model).

The evaluation scores of the XGB machine learning model are very positive and bring high expectations. To solve the information gap of the elderly problem, we can use these effective models to predict the elderly object. Then, we can give some solutions to help them in a society with a strong information technology base. For example, the authorities can give these people introductory training courses on information access skills to limit the information access gap of the elderly.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07041091, 2021S1A5A8062526) and 2022 Development of Open-Lab based on 4P in the Southeast Zone.

References

[1] R. Susło, M. Paplicki, K. Dopierała, and J. Drobnik., Fostering digital literacy in the elderly as a means to secure their health needs and human rights in the reality of the twenty-first century. Family Medicine & Primary Care Review, vol. 20, no. 3, pp. 271-275, 2018. doi: 10.5114/fmpcr.2018.78273.

[2] G. M. Van Jaarsveld, The Effects of COVID-19 Among the Elderly Population: A Case for Closing the Digital Divide. Frontiers in psychiatry vol. 11 577427, 2020. doi: 10.3389/fpsyt.2020.577427.

[3] K. Hatamnezhad, P. Ghafari Ashtiyani, and F. Seyedi, Investigating the Relationship Between Electronic Literacy and Quality of Life of the Elderly in Arak, Iran. Bulletin of Science, Technology & Society, vol. 41, no. 1, Feb. 2021, pp. 3–9. doi: 10.1177/02704676211007360.

[4] A. Powell, A. Bryne, and D. Dailey, The essential Internet: Digital exclusion in low‐income American communities. Policy & Internet, vol. 2, no. 2, pp. 161-192, 2010. doi: 10.2202/1944-2866.1058.

[5] B. Schäffer, The Digital Literacy of Seniors. Research in Comparative and International Education, vol. 2, no. 1, pp. 29–42, 2007. doi: 10.2304/rcie.2007.2.1.29.

[6] D. Castilla, C. Botella, I. Miralles, J. Bretón-López, A. M. Dragomir-Davis, I. Zaragoza, and A. Garcia-Palacios, Teaching digital literacy skills to the elderly using a social network with linear navigation: A case study in a rural area. International Journal of Human-Computer Studies, vol. 118, pp. 24-37, 2018. doi: 10.1016/j.ijhcs.2018.05.009.

[7] W. K. Shin, D. B. Lee, and M. Y. Park, Smartphone Adoption using Smartphone Use and Demographic Characteristics of Elderly. Journal of the Ergonomics Society of Korea, vol. 31, no. 5, pp. 695-704, 2012. doi: 10.5143/jesk.2012.31.5.695.

[8] H. Lee, and S. H. Lee, A study on the relationship between level of digital informatization and satisfaction level of elderly people: Focusing on community, meeting, and community involvement activities. Journal of Digital Convergence, vol. 17, no. 2, pp. 1-7, 2019. doi: 10.14400/JDC.2019.17.2.001.

[9] C. H. Wang, and C. L. Wu, Bridging the digital divide: the smart TV as a platform for digital literacy among the elderly. Behaviour & Information Technology, pp. 1-14, 2021. doi: 10.1080/0144929X.2021.1934732.

[10] S. K. Choi, Current status of digital information gap for women with disabilities from a gender-conscious perspective and ways to support informatization education based on empowerment. Journal of the Korea Institute of Information and Communication Engineering, vol. 24, no. 5, pp. 655-661, 2020. doi:10.6109/JKIICE.2020.24.5.655.

[11] Y. Liu, Y. Wang, and J. Zhang, New Machine Learning Algorithm: Random Forest. In Information Computing and Applications. Springer: Berlin/Heidelberg, Germany, 2012, pp. 246–252.

[12] P. Geurts, D. Ernst, and L. Wehenkel, Extremely randomized trees. Machine Learning, vol. 63, pp. 3-42, 2006. doi: 10.1007/s10994-006-6226-1.

[13] A. O. Alsayed, M. S. M. Rahim, I. AlBidewi, M. Hussain, S. H. Jabeen, N. Alromema, S. Hussain, and M. L. Jibril, Selection of the Right Undergraduate Major by Students Using Supervised Learning Techniques. Applied Sciences, vol. 11, no. 22, pp. 10639, 2021. doi: 10.3390/app112210639.

[14] H. Byeon, Developing a nomogram for predicting the depression of senior citizens living alone while focusing on perceived social support. World journal of psychiatry, vol. 11, no. 12, pp. 1314, 2021. doi: 10.5498/wjp.v11.i12.1314.

[15] N. H. Chowdhury, M. B. I. Reaz, F. Haque, S. Ahmad, S. H. M. Ali, A. A. A Bakar, and M. A. S. Bhuiyan. Performance Analysis of Conventional Machine Learning Algorithms for Identification of Chronic Kidney Disease in Type 1 Diabetes Mellitus Patients. Diagnostics, vol. 11, no. 12, pp. 2267, 2021. doi: 10.3390/diagnostics11122267.

[16] C. Beaulac, and J. S. Rosenthal, Predicting University Students' Academic Success and Major Using Random Forests. Research in Higher Education, vol. 60, pp. 1048-1064, 2019. doi: 10.1007/s11162-019-09546-y.

[17] R. Patil, and S. Tamane, A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes. International Journal of Electrical and Computer Engineering, vol. 8, no. 5, pp. 3966-3975, 2018. doi: 10.11591/ijece.v8i5.pp3966-3975.

[18] A. Asselman, M. Khaldi, and S. Aammou, Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. Interactive Learning Environment, pp, 1-20, 2021. doi: 10.1080/10494820.2021.1928235.

[19] J. Bergstra, D. Yamins, and D. D. Cox, Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. Proceedings of the 30th International Conference on Machine Learning, vol. 28, no. 1, pp. 115-123, 2013.