PPSNV:A novel predictor for pathogenicity of SNV based on ensemble learning

ZHEN XUYANG

Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China

Shanghai Key Laboratory of Psychotic Disorders, Shanghai 200030, China

E-mail: [zhenxuyang@sjtu.edu.cn](mailto:zhenxuyang@sjtu.edu.cn)

LIN GUANNING[[1]](#footnote-1)\*

Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China

Shanghai Key Laboratory of Psychotic Disorders, Shanghai 200030, China

E-mail: [nickgnlin@sjtu.edu.cn](mailto:nickgnlin@sjtu.edu.cn)

With the next-generation sequencing(NGS) technologies developing, numerous genetic data emerges. Nonsynonymous single nucleotide variant(SNV) is a common type of genetic variants which possibly lead to diseases. However, classifying SNVs in benign and pathogenic variants with high confidence remains challenging. Inspired by ensemble learning and a machine learning algorithm, Gradient Boosting Decision Tree(GBDT), we proposed a novel prediction model named as PPSNV to identity the pathogenicity of SNVs. We integrated 14 features to train our model and tested it in two independent datasets. The results showed best performance was achieved by the proposed model compared with four commonly used predictive tools.

**CCS CONCEPTS** • Computing methodologies • Machine learning • Classification and regression trees • Ensemble methods • Boosting

**Keywords:** Single Nucleotide Variant, Pathogenicity prediction, Machine Learning, Gradient Boosting Decision Tree, Ensemble Learning.

1. Introduction

With the next-generation sequencing(NGS) technologies developing, situations become routinely common that doctors conduct genetic variant detection for patients in clinical diagnosis. Providing helpful insight into individual heredity and human diversity, genetic testing is vital for personalized medicine and targeted therapy. Among numerous output of genome sequencing, nonsynonymous single nucleotide variant(SNV) is a common type of genetic variants which probably cause disease, which means aligning SNVs with phenotypes and diseases accurately are necessary. However, identifying the pathogenicity of SNVs by experiment is costly and infeasible and calling for computational methods of pathogenicity prediction of SNVs is always urgent. Concurrently, actual consequence of genetic variants in vivo veils behind complicated biological mechanisms through many pathways. Hench, discriminating SNVs between benign and pathogenic variants with high confidence remains challenging.

Based on different roadmaps of genetic central dogma, many tools are developed to predict the pathogenicity of genetic variants. Some tools utilizes the conservation of amino acid or nucleotide to assess the damage when the reference allele is substituted, and several predictive tools focus on the alteration of molecular mechanisms, such as the structure or functional change of protein. Trained with different algorithms and individually collected data, different tools may provide different results for the same genetic variant. However, most tools are still handy references for the assessment of genetic variants. Extra information of SNVs that may represent deep predictive features in machine learning are refined during the prediction of developed tools in silicon, which probably facilitates related research.

As a strategy of machine learning, ensemble learning is proved to acquire advantages from its base learners. Through the combination of various existing predictors, better prediction results are obtained without additional primary features. Inspired by it and integrating several existing predictors, we propose PPSNV, a novel predictor for pathogenicity of SNV based on ensemble learning in this study. We implement PPSNV by Gradient Boosting Decision Tree(GBDT), a machine learning algorithm that perform well in classification and regression task. With fine parameter tuning, we achieve our predictive model and test it with two independent datasets and compare it with four ensemble predictors, PPSNV outperforms all compared predictors overall.

1. METHODS
   1. Gradient Boosting Decision Tree(GBDT)

Utilizing a boosting strategy, Gradient Boosting Decision Tree(GBDT) is a method which achieves state-of-the-art performances among other machine learning methods in many tasks of classification and regression. Previous research works have used GBDT to perform prediction in biological problem and achieved best performance. Give its superior classification performance, we build our pathogenic predictors based on GBDT. Especially, we use LightGBM, a novel implementation of GBDT which succeeds in dealing with large number of data instances and sparse or not features, to finish our model. In terms of categorical features and numerical features of our SNVs, LightGBM is suitable for the classification.

* 1. Datasets

We use the dataset provided by a previous work, containing 5 subsets and 88,332 variants which are collected from publicly available and commonly used benchmark datasets. After removing overlapping variants between five subsets, *HumVar, VariBenchSelected, predictSNPSelected* are merged into a new dataset which we use to train our model. The merged dataset, consisting of 66577 SNVs, is split into split into a training set and a validation set in an 7:3 ratio.

As for *ExoVar and SwissVarSelected,* we leave them as two independent test datasets based on which we compare our proposed predictors with others. Annovar is used to annotate all variants in our datasets, by which we obtained input features of our model and output of the comparators. Especially, we removed a variant in the test datasets if its output of the comparators is missing.

The details of our datasets are listed in Table 1.

Table 1. The Datasets used in this study

| Dataset | Benign(B) | Pathogenic(P) | Total |
| --- | --- | --- | --- |
| Training | 21,740 | 24,863 | 46,603 |
| Vaildation | 9,455 | 10,519 | 19,974 |
| ExoVar | 568 | 4,984 | 5,552 |
| SwissVarSelected | 5,004 | 4,356 | 9,360 |

* 1. Training

The classification performance of GBDT depends on how the decision trees are constructed during training. We used GridSearch provided by Scikit-learn to tune the parameters. The core parameters, such as learning rate, max depth of the decision trees and number of stacked estimators, were optimized to achieve best performance. We trained our model by maximizing the AUC score for binary pathogenicity classification. The initial outcome of a SNV from our predictor is a probability score ranging between 0 and 1 which indicates pathogenicity of the SNV. We use the threshold of 0.5 as a cut-off to classify the SNV into a benign or pathogenic variant(pathogenic if the score > 0.5).

* 1. Features

We include 14 features in total to train PPSN. The first part of our features is four primary features of SNV(chromosome, location, reference amino acid and alternate amino acid) and the ohter part is ten pathogenicity predictions scores(SIFT, Polyphen2\_HDIV, Polyphen2\_HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, RadialSVM, VEST3, CADD) obtained from existing tools. Missing features are not imputed as LightGBM enables the missing value handle. Actually, permitting missing value in the input features is friendly for practical application considering that many SNVs may not get valid result from some existing prediction tools.

Table 2. The Features used in this study

| Feature | Description |
| --- | --- |
| Chr | The chromosome |
| Vaildation | 9,455 |
| ExoVar | 568 |
| SwissVarSelected | 5,004 |

* 1. Testing

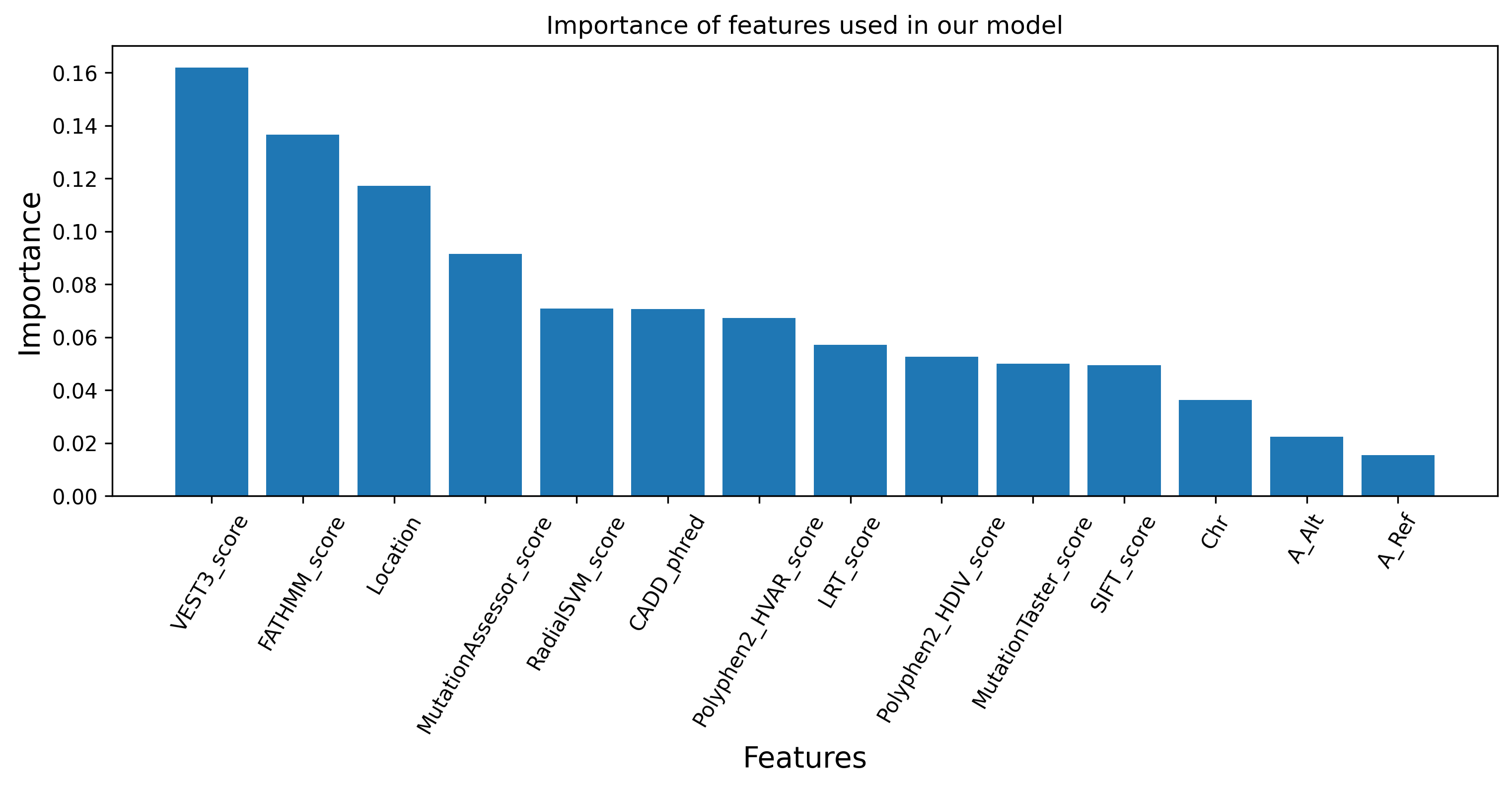
We deploy two independent test datasets to evaluate PPSNV’s performance and compare it with other predictors. In order to avoid basis in comparation, we assure that test data did not overlap with our training data. We use trained model to predict each SNVs in test datasets. Given that several tools return pathogenicity score that are out of range between 0 to 1, we normalized its output to [0, 1] for test SNVs.

* 1. Comparators

PPSNV are compared with four predictions tools that are commonly used and readily implemented：M-CAP, REVEL, MVP, DANN which are all prediction tools based on ensemble learning. The predictive scores for test data of these tools are annotated by Annovar. Each SNVs in our test datasets obtain all score of comparators without missing value. Several compared tools return pathogenicity scores instead of binary classification results. The optimal thresholds recommended to discriminate SNVs pathogenic or benign are not explicit, which may be misleading if we used binary metrics such as precision or recall to evaluate classification performance. Hence, we apply receiver operating characteristic(ROC) curves, the area under the ROC curve(AUC) and precision-recall (PR) curve to assess the overall performance of five predictors including the proposed.

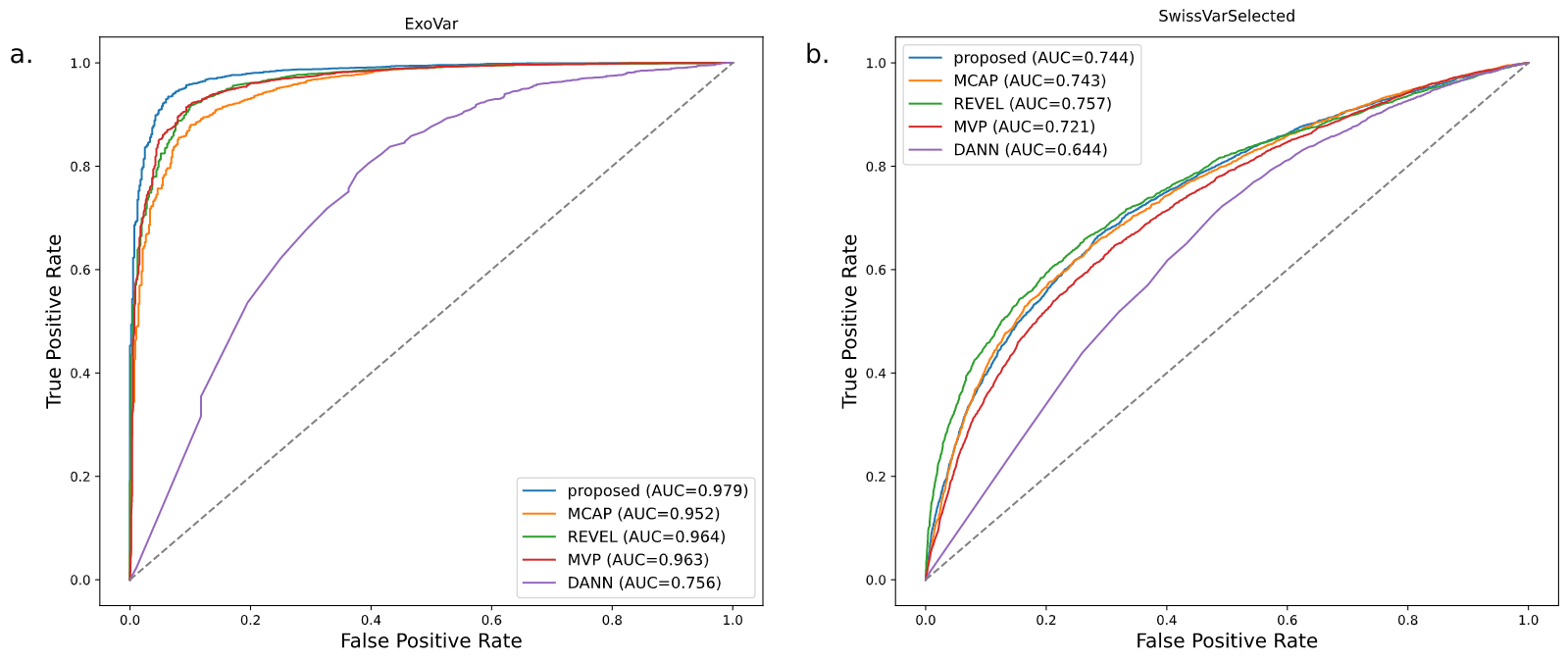
1. RESULT
   1. Characterization of Features

PPSNV ensemble the pathogenicity predictions of 10 individual tools as features besides 4 primary features of SNV. Importance of features in PPSNP represents information gain of features in the algorithm. With higher importance, features contribute more to the final classification result. We normalized the importance of features(Figure 1) and the top five important features in our model are VEST3 score, FATHMM score, Location, MutationAssessor score and RadialSVM score. Compared with primary features, features from ensemble tools earn much higher importance which indicates the prediction scores of existing tools represents more intrinsic information related with underlying biological mechanisms.



* 1. Performance of PPSNV compared with other methods

The PPSNV performs well in the test with an AUC of 0.745, which surpasses M-CAP, MVP and DANN. However, its performance is slightly worse than the performance of REVEL. Considering that REVEL ensembles 18 individual pathogenicity prediction score, PPSNV is a lightweight model.



1. DISCUSSION

Aiming to predict the pathogenicity of genetic variants accurately, there are many methods which base on different biological mechanism.

ACKNOWLEDGMENTS

This work was supported by grants from the National Natural Science Foundation of China (nos. 81971292), the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (no. 1610000043), and the Innovation Research Plan supported by Shanghai Municipal Education Commission (ZXWF082101).

REFERENCES

1. Atul Adya, Paramvir Bahl, Jitendra Padhye, Alec Wolman, and Lidong Zhou. 2004. A multi-radio unification protocol for IEEE 802.11 wireless networks. In Proceedings of the IEEE 1st International Conference on Broadnets Networks (BroadNets’04) . IEEE, Los Alamitos, CA, 210–217. https://doi.org/10.1109/BROADNETS.2004.8
2. Sam Anzaroot and Andrew McCallum. 2013. UMass Citation Field Extraction Dataset. Retrieved May 27, 2019 from <http://www.iesl.cs.umass.edu/data/data-umasscitationfield>
3. Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24, 6 (June 1981), 381–395. https://doi.org/10.1145/358669.358692
4. Chelsea Finn. 2018. Learning to Learn with Gradients. PhD Thesis, EECS Department, University of Berkeley.
5. Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. J. ACM 46, 5 (September 1999), 604–632. https://doi.org/10.1145/324133.324140
6. Matthew Van Gundy, Davide Balzarotti, and Giovanni Vigna. 2007. Catch me, if you can: Evading network signatures with web-based polymorphic worms. In Proceedings of the first USENIX workshop on Offensive Technologies (WOOT ’07) . USENIX Association, Berkley, CA, Article 7, 9 pages.
7. James W. Demmel, Yozo Hida, William Kahan, Xiaoye S. Li, Soni Mukherjee, and Jason Riedy. 2005. Error Bounds from Extra Precise Iterative Refinement. Technical Report No. UCB/CSD-04-1344. University of California, Berkeley.
8. David Harel. 1979. First-Order Dynamic Logic. Lecture Notes in Computer Science, Vol. 68. Springer-Verlag, New York, NY. <https://doi.org/10.1007/3-540-09237-4>
9. Jason Jerald. 2015. The VR Book: Human-Centered Design for Virtual Reality. Association for Computing Machinery and Morgan & Claypool.
10. Prokop, Emily. 2018. The Story Behind. Mango Publishing Group. Florida, USA.
11. R Core Team. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
12. Brian K. Reid. 1980. A high-level approach to computer document formatting. In Proceedings of the 7th Annual Symposium on Principles of Programming Languages. ACM, New York, 24–31. <https://doi.org/10.1145/567446.567449>
13. John R. Smith and Shih-Fu Chang. 1997. VisualSEEk: a fully automated content-based image query system. In Proceedings of the fourth ACM international conference on Multimedia (MULTIMEDIA ’96). Association for Computing Machinery, New York, NY, USA, 87–98. https://doi.org/10.1145/244130.244151
14. TUG 2017. Institutional members of the LaTeX Users Group. Retrieved May 27, 2017 from <http://wwtug.org/instmem.html>
15. Alper Yilmaz, Omar Javed, and Mubarak Shah. 2006. Object tracking: A survey. ACM Comput. Surv. 38, 4 (December 2006), 13–es. https://doi.org/10.1145/1177352.1177355

Abdelghaffar, H., Kamel, S., and Duquenoy, P. (2010). “Studying E-Government Trust in Developing Nations: Case of University and Colleges Admissions and Services in Egypt”. Proceedings of the International Information Management Association Conference, Utrecht, The Netherlands.

Abdelsalam, H., Reddick, C., Gamal, S., and Al-shaar, A. (2013). “Social Media in Egyptian Government Websites: Presence, Usage, and Effectiveness”. Government Information Quarterly, 30(4): 406-416.

Baldassare, M. (2000). “California in the New Millennium: The Changing Social and Political Landscape”. Berkeley: University of California Press.

Bertot, J. and Grimes, J. (2012). “Promoting Transparency and Accountability through ICTs, Social Media, and Collaborative E-Government”. Transforming Government: People, Process and

A

Non pulvinar neque laoreet suspendisse interdum consectetur libero. Facilisis leo vel fringilla est ullamcorper eget nulla facilisi. Ipsum dolor sit amet consectetur adipiscing elit pellentesque. Risus quis varius quam quisque id. Bibendum arcu vitae elementum curabitur vitae. Vitae et leo duis ut diam quam nulla. Orci eu lobortis elementum nibh tellus molestie nunc non blandit. Arcu odio ut sem nulla pharetra diam sit amet. Quis vel eros donec ac odio. Est lorem ipsum dolor sit amet consectetur adipiscing

Authors’ Information

|  |  |  |  |
| --- | --- | --- | --- |
| Your Name | Title\* | Research Field | Personal website |
| First Author |  |  |  |
| Second Author |  |  |  |
| Third Author |  |  |  |

\*This form helps us to understand your paper better, and the form itself will not be published.

\*Title can be chosen from: Master Student, PhD Candidate, Doctor, Assistant Professor, Lecturer, Senior Lecturer, Associate Professor, Full Professor, Researcher, Senior Researcher

1. \* Corresponding author [↑](#footnote-ref-1)