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

This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word

FIRST AUTHOR'S NAME, INITIALS, AND LAST NAME[[1]](#footnote-1)\*

First author's affiliation, an Institution with a very long name; Email

SECOND AUTHOR'S NAME, INITIALS, AND LAST NAME

Second author's affiliation, possibly the same institution; Email

THIRD AUTHOR'S NAME, INITIALS, AND LAST NAME

Third author's affiliation, possibly the same institution; Email

Although there is no distinctive header, this is the abstract. This submission template allows authors to submit their papers for review to an ACM Conference or Journal without any output design specifications incorporated at this point in the process. The ACM manuscript template is a single column document that allows authors to type their content into the pre-existing set of paragraph formatting styles applied to the sample placeholder text here. Throughout the document you will find further instructions on how to format your text.

**CCS CONCEPTS** • Insert your first CCS term here • Insert your second CCS term here • Insert your third CCS term here

**Additional Keywords and Phrases:** Insert comma delimited author-supplied keyword list, Keyword number 2, Keyword number 3, Keyword number 4

ACM Reference Format:

First Author’s Name, Initials, and Last Name, Second Author’s Name, Initials, and Last Name, and Third Author’s Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. In Woodstock ’18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA, 10 pages. NOTE: This block will be automatically generated when manuscripts are processed after acceptance.

1. Introduction

With the next-generation sequencing (NGS) technologies developing, situations become common that doctors conduct genetic testing for their patients in clinical diagnosis. Linking disease with responsible genetic variants correctly is vital for personalized medicine and treatment. Among numerous output of genome sequencing, nonsynonymous single nucleotide variant(SNV) is a common type of genetic variants which probably cause disease. However, to identify the pathogenicity of SNVs by experiment is costly and time-consuming and call for computational methods for classification of SNVs is always urgent. But specific consequence of genetic variants in vivo veils behind complicated biological mechanisms in many pathways, which means that discriminating SNVs between benign and pathogenic with high confidence remains challenging.

Based on different properties through roadmaps of genetic central dogma, many tools are developed to predict the pathogenicity of genetic variants. A part of these tools utilizes the conservation of amino acid or nucleotide to assess the damage when the reference allele is substituted, and several other predictive tools focus on the alteration of molecular mechanisms, such as the structure change of protein. Though different tools may provide different results for the same genetic variant, most tools are handy references for the assessment of genetic variants. Underlaying information of SNVs, which we call as deep predictive features in machine learning, are refined during the prediction of individual tools in silicon.

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 probably 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. We test it on an independent dataset and compare it with four ensemble predictors, PPSNV performs excellently with less integrated predictors.

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

Based on boosting strategy, Gradient Boosting Decision Tree(GBDT) is a method which outstands among other machine learning methods in the tasks of classification and regression. Some previous research works use GBDT to perform prediction in biological tasks and achieve best performance(Liao, Huang et al. 2016). Specially, GBDT is suitable for tabular and dense input data space which our data exists. Inspired by what mentioned above, we propose a framework based on GBDT to predict the pathogenicity of SNP.

We use LightGBM, an implementation and optimization of GBDT, to build our model.

* 1. Datasets

We use the dataset provided of a previous work, containing 5 subsets and 45081 variants which are collected from publicly available and commonly used benchmark datasets. All variants in the dataset are labeled and the variants overlapping with CADD training data are removed. After removing overlapping variants between 5 subsets, *HumVar, ExoVar, VariBenchSelected, predictSNPSelected* are merged into a new dataset which we use to train our model. As for *SwissVarSelected,* we leave it as an independent test dataset based on which we compare our proposed framework with other methods.

The details of our datasets are listed below.

* 1. Training

Since PPSNV is based on the open-source software library, LightGBM which provides python implementation, we tune the hyperparameters of our classifier, such as the learning rate and the max depth of trees using Scikit-learn(Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.) . Besides, we adopt an early stopping strategy to avoid over-fitting. Our proposed model is trained by maximizing the AUC for binary pathogenicity classification.

* 1. Features

We include 16 features in total to train PPSNV. Predictive features included can be divided into two parts, one of which is 6 basic features of SNV(chromosome, location, reference allele, alternate allele, reference amino acid, alternate amino acid) and others are pathogenicity predictions scores from 10 existing tools(SIFT, Polyphen2\_HDIV, Polyphen2\_HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, RadialSVM, VEST3, CADD). All features are obtained from Annovar. Missing features are not imputed since our model is rubost to handle some missing values in input features. Besides, tolerating missing input is suitable for practical prediction because many SNVs are annotated by existing prediction tools but some tools do not return valid result.

* 1. Testing
  2. Comparators

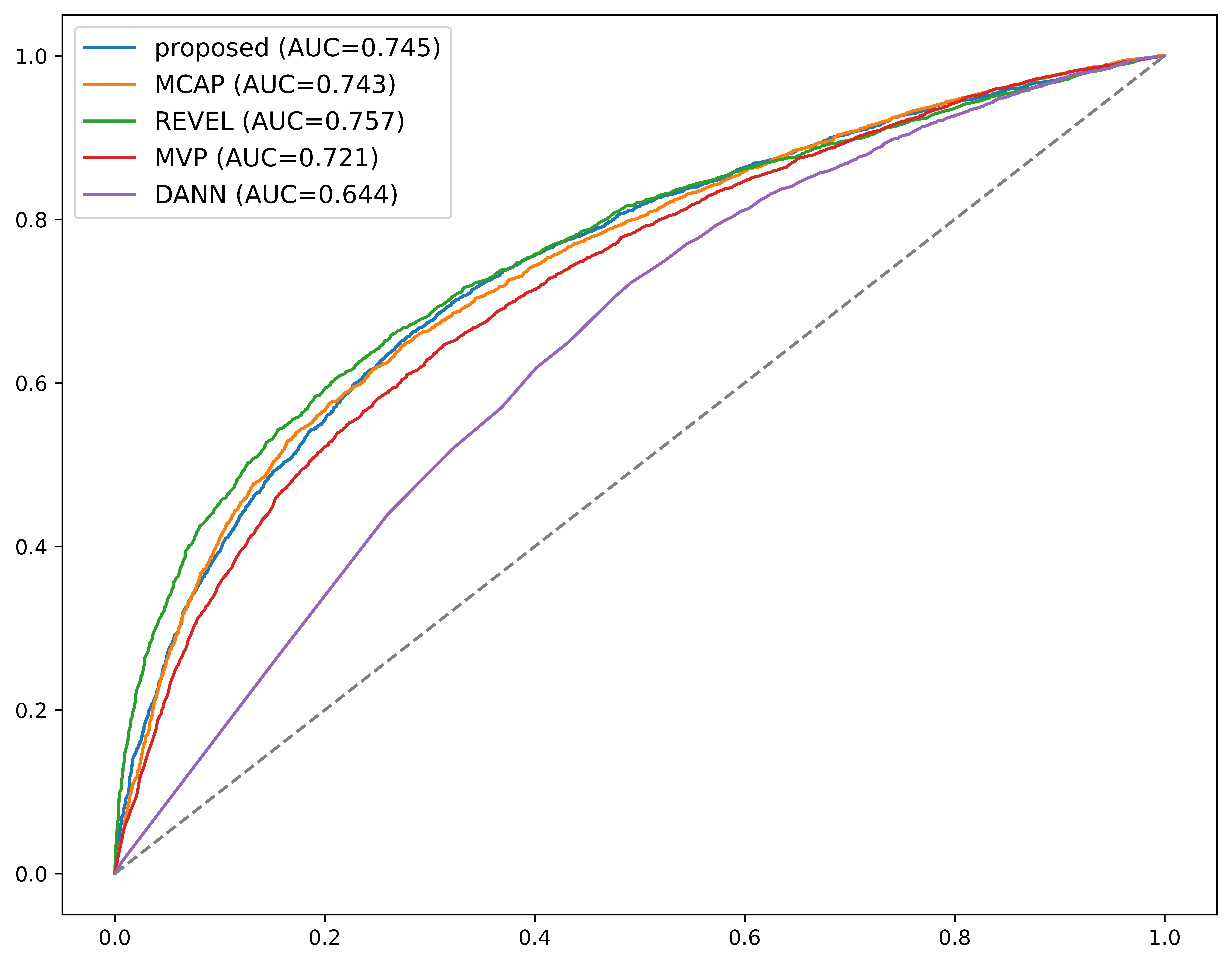
PPSNV are compared with four predictions tools that are commonly used and readily implemented：M-CAP, REVEL, MVP, DANN which are all ensemble prediction tools. The results of these tools are obtained by Annovar and the SNVSs in our test dataset with missing values are removed. Considering that several compared tools return pathogenic scores instead of binary classification results and the thresholds recommended to decide a SNV pathogenic or not may be not optimal in our research, 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 six predictors including the proposed accurately.

1. RESULT
   1. Characterization of Features

The PPSNV ensemble the pathogenic predictions of 10 individual tools as features besides 6 basic properties of SNV. The importance of a feature in PPSNP reflects how the feature contributes to the final prediction result. The top five important features in our model are VEST3, FATHMM, CADD, MutationAssessor, RadialSVM. Compared with the basic features, features from ensemble tools earn much higher importance which indicates the prediction scores of existing tools represents more intrinsic information.

* 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.

1. Citing Related Work

This section cites a variety of journal [5, 15], conference [1, 6, 8, 12, 13], and magazine [3] articles to illustrate how they appear in the references section. It also cites books [9, 10], a technical report [7], a PhD dissertation [4], an online reference [14], a software artifact [11], and a dataset [2].

ACKNOWLEDGMENTS

Acknowledgments are placed before the references. Add information about grants, awards, or other types of funding that you have received to support your research. Author can capture the **grant sponsor information**, by selecting the grant sponsor text and apply style ‘GrantSponsor’. After this, select grant no and apply ‘GrantNumber’ from style panel. Example of Grant sponsor: Competitive Research Programme and example of Grant no: CRP 10-2012-03.

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  APPENDICES

In the appendix section, three levels of Appendix headings are available.

A.1 General Guidelines (AppendixH2)

1. Save as you go and backup your file regularly.
2. Do not work on files that are saved in a cloud directory. To avoid problems such as MS Word crashing, please only work on files that are saved locally on your machine.
3. Equations should be created with the built-in Microsoft® Equation Editor included with your version of Word. (Please check the compatibility at <http://tinyurl.com/lzny753> for using MathType.)
4. Please save all files in DOCX format, as the DOC format is only supported for the Mac 2011 version.
5. Tables should be created with Word’s “Insert Table” tool and placed within your document. (Tables created with spaces or tabs will have problems being properly typeset. To ensure your table is published correctly, Word’s table tool must be used.)
6. Do not copy-and-paste elements into the submission document from Excel such as charts and tables.
7. Footnotes should be inserted using Word’s “Insert Footnote” feature.
8. Do not use Word’s “Insert Shape” function to create diagrams, etc.
9. Do not have references appear in a table/cells format as it will produce an error during the layout generation process.
10. MS Word does not consistently allow the original formatting to be modified in the text. In these cases, it is best to copy all the document’s text from the specific file and paste into a new MS Word document and then save it.
11. At times there are font problems such as “odd” stuff/junk characters that appear in the text, usually in the references. This can be caused by a variety of reasons such as copying-and-pasting from another file, file transfers, etc. Please review your text prior to submission to make sure it reads correctly.

A.1.1 Preparing Graphics (AppendixH3)

1. Accepted image file formats: TIFF (.tif), JPEG (.jpg).
2. Scalable vector formats (i.e., SVG, EPS and PS) are greatly preferred.
3. Application files (e.g., Corel Draw, MS Word, MS Excel, PPT, etc.) are NOT recommended.
4. Images created in Microsoft Word using text-box, shapes, clip-art are NOT recommended.
5. IMPORTANT: All fonts must be embedded in your figure files.
6. Set the correct orientation for each graphics file.

A.2 Placeholder Text

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Vulputate sapien nec sagittis aliquam. Malesuada fames ac turpis egestas sed tempus urna. Posuere sollicitudin aliquam ultrices sagittis orci. Consequat id porta nibh venenatis cras sed felis eget. Pellentesque eu tincidunt tortor aliquam nulla facilisi cras fermentum odio. Tincidunt nunc pulvinar sapien et ligula ullamcorper malesuada proin. Tincidunt lobortis feugiat vivamus at augue. Eget nunc lobortis mattis aliquam faucibus. Egestas diam in arcu cursus euismod quis.

Erat pellentesque adipiscing commodo elit at imperdiet. In hac habitasse platea dictumst quisque sagittis purus. At lectus urna duis convallis. Eu mi bibendum neque egestas congue. Est ullamcorper eget nulla facilisi etiam dignissim diam. Sed ullamcorper morbi tincidunt ornare massa eget. Aenean vel elit scelerisque mauris pellentesque. Ullamcorper dignissim cras tincidunt lobortis feugiat vivamus. Cras fermentum odio eu feugiat pretium nibh. Congue eu consequat ac felis donec et odio pellentesque diam. Velit sed ullamcorper morbi tincidunt ornare massa eget egestas. In metus vulputate eu scelerisque felis imperdiet proin fermentum leo. Nulla malesuada pellentesque elit eget gravida cum.

Nullam ac tortor vitae purus faucibus ornare suspendisse. Libero enim sed faucibus turpis in eu mi bibendum neque. Sodales ut etiam sit amet nisl purus. Egestas diam in arcu cursus. Aliquet porttitor lacus luctus accumsan tortor. Pharetra magna ac placerat vestibulum lectus. Sit amet mauris commodo quis imperdiet massa tincidunt. In nisl nisi scelerisque eu ultrices vitae auctor. Nisi vitae suscipit tellus mauris a diam. Dui vivamus arcu felis bibendum ut tristique. Laoreet suspendisse interdum consectetur libero id.

Enim eu turpis egestas pretium. Nulla aliquet enim tortor at auctor urna. Id aliquet risus feugiat in. Non enim praesent elementum facilisis leo. Integer feugiat scelerisque varius morbi enim nunc faucibus. Egestas dui id ornare arcu odio ut sem nulla pharetra. Montes nascetur ridiculus mus mauris. Orci dapibus ultrices in iaculis. Enim sed faucibus turpis in eu mi bibendum neque. Faucibus pulvinar elementum integer enim neque volutpat ac tincidunt vitae. Et ultrices neque ornare aenean euismod elementum. Et pharetra pharetra massa massa ultricies mi quis hendrerit dolor. Tempus iaculis urna id volutpat lacus laoreet non curabitur gravida. Est velit egestas dui id ornare arcu odio. Eu facilisis sed odio morbi quis commodo odio. Lectus vestibulum mattis ullamcorper velit sed ullamcorper morbi tincidunt.

Eu non diam phasellus vestibulum lorem sed risus ultricies. Convallis aenean et tortor at risus viverra adipiscing at. Mauris pellentesque pulvinar pellentesque habitant morbi. Elementum sagittis vitae et leo duis. Massa enim nec dui nunc. Nisl tincidunt eget nullam non nisi est sit amet. Amet nisl purus in mollis nunc sed id semper. Fermentum leo vel orci porta non pulvinar neque laoreet suspendisse. Diam vel quam elementum pulvinar etiam non quam. Sagittis orci a scelerisque purus semper eget. Aliquet porttitor lacus luctus accumsan tortor. Integer vitae justo eget magna fermentum iaculis eu non diam. Egestas pretium aenean pharetra magna ac. Cursus metus aliquam eleifend mi in nulla. Cursus mattis molestie a iaculis at erat pellentesque adipiscing. Pulvinar pellentesque habitant morbi tristique senectus. Gravida cum sociis natoque penatibus et magnis dis parturient montes. In aliquam sem fringilla ut. Ut consequat semper viverra nam libero justo laoreet. Pellentesque diam volutpat commodo sed egestas.

Ornare arcu odio ut sem nulla pharetra diam. Ut enim blandit volutpat maecenas volutpat blandit aliquam. Tempus iaculis urna id volutpat lacus. Nascetur ridiculus mus mauris vitae. Venenatis cras sed felis eget velit aliquet sagittis id. Laoreet non curabitur gravida arcu ac tortor dignissim convallis aenean. Maecenas ultricies mi eget mauris pharetra et ultrices neque ornare. Egestas purus viverra accumsan in nisl nisi scelerisque eu ultrices. Tempus urna et pharetra pharetra massa massa. Pulvinar neque laoreet suspendisse interdum consectetur libero id.

Nisl rhoncus mattis rhoncus urna neque viverra justo nec ultrices. Morbi quis commodo odio aenean sed adipiscing diam donec. Neque gravida in fermentum et. Scelerisque purus semper eget duis at tellus. Volutpat blandit aliquam etiam erat velit scelerisque in dictum non. Odio ut sem nulla pharetra diam sit. Sed pulvinar proin gravida hendrerit lectus a. Diam ut venenatis tellus in metus vulputate eu scelerisque. Id semper risus in hendrerit. Vel quam elementum pulvinar etiam. Amet aliquam id diam maecenas ultricies mi. Auctor elit sed vulputate mi sit amet. Orci dapibus ultrices in iaculis nunc. Sed vulputate odio ut enim blandit volutpat maecenas volutpat. Auctor urna nunc id cursus metus. Integer enim neque volutpat ac tincidunt vitae.

Scelerisque in dictum non consectetur a erat. Vel risus commodo viverra maecenas accumsan lacus vel facilisis volutpat. Dignissim sodales ut eu sem integer vitae justo eget magna. Nunc non blandit massa enim nec dui nunc mattis enim. Sed vulputate odio ut enim blandit volutpat maecenas. Ante in nibh mauris cursus. Donec pretium vulputate sapien nec sagittis aliquam malesuada. Eu volutpat odio facilisis mauris sit amet massa. Blandit turpis cursus in hac habitasse platea dictumst quisque. Donec enim diam vulputate ut pharetra sit.

Magna fringilla urna porttitor rhoncus dolor purus non. Fames ac turpis egestas integer eget. Mattis rhoncus urna neque viverra. Laoreet sit amet cursus sit amet dictum sit amet. Vel pretium lectus quam id leo in vitae turpis massa. Euismod lacinia at quis risus sed vulputate odio ut. Lorem dolor sed viverra ipsum. Viverra justo nec ultrices dui sapien. Aliquam nulla facilisi cras fermentum odio eu feugiat pretium. Adipiscing commodo elit at imperdiet dui accumsan sit amet nulla. Morbi leo urna molestie at elementum eu facilisis sed. Habitant morbi tristique senectus et netus et malesuada. Viverra ipsum nunc aliquet bibendum enim. Integer vitae justo eget magna fermentum. Tincidunt id aliquet risus feugiat. Mauris ultrices eros in cursus turpis. Amet venenatis urna cursus eget nunc. Nisl nisi scelerisque eu ultrices vitae.

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

Liao, Z., et al. (2016). "<i>In Silico</i> Prediction of Gamma-Aminobutyric Acid Type-A Receptors Using Novel Machine-Learning-Based SVM and GBDT Approaches." BioMed Research International **2016**: 2375268.

Gamma-aminobutyric acid type-A receptors (<inline-formula><mml:math xmlns:mml="<http://www.w3.org/1998/Math/MathML>" id="M1"><mml:msub><mml:mrow><mml:mi mathvariant="normal">G</mml:mi><mml:mi mathvariant="normal">A</mml:mi><mml:mi mathvariant="normal">B</mml:mi><mml:mi mathvariant="normal">A</mml:mi></mml:mrow><mml:mrow><mml:mi mathvariant="normal">A</mml:mi></mml:mrow></mml:msub><mml:mi mathvariant="normal">R</mml:mi></mml:math></inline-formula>s) belong to multisubunit membrane spanning ligand-gated ion channels (LGICs) which act as the principal mediators of rapid inhibitory synaptic transmission in the human brain. Therefore, the category prediction of <inline-formula><mml:math xmlns:mml="<http://www.w3.org/1998/Math/MathML>" id="M2"><mml:msub><mml:mrow><mml:mi mathvariant="normal">G</mml:mi><mml:mi mathvariant="normal">A</mml:mi><mml:mi mathvariant="normal">B</mml:mi><mml:mi mathvariant="normal">A</mml:mi></mml:mrow><mml:mrow><mml:mi mathvariant="normal">A</mml:mi></mml:mrow></mml:msub><mml:mi mathvariant="normal">R</mml:mi></mml:math></inline-formula>s just from the protein amino acid sequence would be very helpful for the recognition and research of novel receptors. Based on the proteins&#x2019; physicochemical properties, amino acids composition and position, a <inline-formula><mml:math xmlns:mml="<http://www.w3.org/1998/Math/MathML>" id="M3"><mml:msub><mml:mrow><mml:mi mathvariant="normal">G</mml:mi><mml:mi mathvariant="normal">A</mml:mi><mml:mi mathvariant="normal">B</mml:mi><mml:mi mathvariant="normal">A</mml:mi></mml:mrow><mml:mrow><mml:mi mathvariant="normal">A</mml:mi></mml:mrow></mml:msub><mml:mi mathvariant="normal">R</mml:mi></mml:math></inline-formula> classifier was first constructed using a 188-dimensional (188D) algorithm at 90&#x25; cd-hit identity and compared with pseudo-amino acid composition (PseAAC) and ProtrWeb web-based algorithms for human <inline-formula><mml:math xmlns:mml="<http://www.w3.org/1998/Math/MathML>" id="M4"><mml:msub><mml:mrow><mml:mi mathvariant="normal">G</mml:mi><mml:mi mathvariant="normal">A</mml:mi><mml:mi mathvariant="normal">B</mml:mi><mml:mi mathvariant="normal">A</mml:mi></mml:mrow><mml:mrow><mml:mi mathvariant="normal">A</mml:mi></mml:mrow></mml:msub><mml:mi mathvariant="normal">R</mml:mi></mml:math></inline-formula> proteins. Then, four classifiers including gradient boosting decision tree (GBDT), random forest (RF), a library for support vector machine (libSVM), and k-nearest neighbor (<inline-formula><mml:math xmlns:mml="<http://www.w3.org/1998/Math/MathML>" id="M5"><mml:mrow><mml:mi>k</mml:mi></mml:mrow></mml:math></inline-formula>-NN) were compared on the dataset at cd-hit 40&#x25; low identity. This work obtained the highest correctly classified rate at 96.8&#x25; and the highest specificity at 99.29&#x25;. But the values of sensitivity, accuracy, and Matthew&#x2019;s correlation coefficient were a little lower than those of PseAAC and ProtrWeb; GBDT and libSVM can make a little better performance than RF and <inline-formula><mml:math xmlns:mml="<http://www.w3.org/1998/Math/MathML>" id="M6"><mml:mrow><mml:mi>k</mml:mi></mml:mrow></mml:math></inline-formula>-NN at the second dataset. In conclusion, a <inline-formula><mml:math xmlns:mml="<http://www.w3.org/1998/Math/MathML>" id="M7"><mml:msub><mml:mrow><mml:mi mathvariant="normal">G</mml:mi><mml:mi mathvariant="normal">A</mml:mi><mml:mi mathvariant="normal">B</mml:mi><mml:mi mathvariant="normal">A</mml:mi></mml:mrow><mml:mrow><mml:mi mathvariant="normal">A</mml:mi></mml:mrow></mml:msub><mml:mi mathvariant="normal">R</mml:mi></mml:math></inline-formula> classifier was successfully constructed using only the protein sequence information.

1. \* Place the footnote text for the author (if applicable) here. [↑](#footnote-ref-1)