Studi Komparasi Performa Algoritma *Supervised-Learning* dalam Klasifikasi Multidataset

I Gusti Ngurah Ryo Aditarta
Fakultas Ilmu Komputer
Universitas Brawijaya
Malang, Indonesia
ryoaditarta@student.ub.ac.id

2nd Dian Pandu Syahfitra
Fakultas Ilmu Komputer
Universitas Brawijaya
Malang, Indonesia
dianpandu@student.ub.ac.id

3rd Rayhan Egar Sadtya Nugraha
Fakultas Ilmu Komputer
Universitas Brawijaya
Malang, Indonesia
rayhanegar@student.ub.ac.id

4th Ahmad Zaki
Fakultas Ilmu Komputer
Universitas Brawijaya
Malang, Indonesia
ahmadzaki12@student.ub.ac.id

5th Muhammad Arya Ghifari
Fakultas Ilmu Komputer
Universitas Brawijaya
Malang, Indonesia
aryaghifary07@student.ub.ac.id

6th Arion Syemael Siahaan
Fakultas Ilmu Komputer
Universitas Brawijaya
Malang, Indonesia
siahaanarion@student.ub.ac.id

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

This document is a model and instructions for LaTeX. Please observe the conference page limits.

II. Algoritma Klasifikasi Supervised Learning

A. K-Nearest Neighbors (KNN)

Algoritma klasifikasi *nearest neighbors* seperti KNN merupakan salah satu algoritma dasar nonparametrik dalam pembelajaran mesin. Hal ini didasarkan pada *rationale* di mana fitur yang digunakan untuk mendeskripsikan label sebuah *domain point* memiliki relevansi dengan *domain point* lain dalam sisi *proximity* [4], [5]. Dalam implementasinya, KNN banyak digunakan dalam permasalahan klasifikasi dengan domain pengetahuan yang terdefinisi dan diketahui dengan baik, seperti klasifikasi tumor otak [6], klasifikasi tingkat keparahan Covid-19 [7], maupun prognosis kanker [8].

Secara prinsip, nilai parameter k yang merepresentasikan banyaknya *domain point* yang digunakan dalam penentuan label serta metode perhitungan jarak merupakan konsiderasi utama. Nilai k yang terlalu kecil memberikan model yang kompleks dan kurang mampu mengakomodasi *unseen data*, sedangkan nilai k yang terlalu besar memberikan model yang terlalu sederhana untuk secara akurat mengklasifikasikan suatu *domain point* [9]. Beberapa teknik, seperti *normalized class coherence*, *change-based KNN*, *variable selection* dan *weighting*, maupun normalisasi L1 dan LPP dapat digunakan untuk memberikan estimasi yang lebih baik [10]–[12]. Selain itu, pilihan perhitungan jarak antara domain point dalam proses pembelajaran model seperti Euclidean "(1)", Manhattan "(2)", maupun Minkowski "(3)" untuk pemetaan label kelas juga perlu untuk diperhatikan.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \tag{1}$$

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \tag{2}$$

$$d(x,y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{\frac{1}{4}} \tag{3}$$

B. Ridge Classifier (RC)

Algoritma klasifikasi *Ridge Classifier* (RC) dikembangkan dari algoritma *Ridge Regression* yang mengkombinasikan *learning rule Regularized Learning Minimization* (RLM) dengan regresi *linear ordinary least squares* [1]. Algoritma RC melakukan *class labelling* berdasarkan tanda/*sign* dari suatu *data point* (positif atau negatif). Penggunaan fungsi regularisasi, seperti regularisasi L2/Tikhonov "(4)" "(5)" pada algoritma klasifikasi memberikan model yang lebih "stabil" dan mencegah terjadinya *overfitting* [4], [5]. Dibandingkan dengan regresi linear, algoritma RC akan meminimalisasi nilai koefisien w untuk masing-masing fitur, memberikan model dengan kemampuan generalisasi yang baik terhadap *unseen data*. Dengan demikian, fungsi yang digunakan dalam RC dapat didefinisikan secara formal pada "(6)".

$$||w|| = \sqrt{\sum_{i=1}^{d} w_i^2} \tag{4}$$

$$\arg min_w(L_s(w) + \lambda||w||^2) \tag{5}$$

$$\arg min_{w \in R^d}(\lambda||w||^2 + \frac{1}{m}\sum_{i=1}^{m}(< w, x_i > -y_i)^2) \quad (6)$$

Dalam melakukan proses pembelajaran model, parameter regularisasi menjadi konsiderasi utama. Semakin kecil nilai, nilai koefisien w akan semakin kecil, memberikan model dengan generalisasi yang baik namun dengan potensi *underfitting*. Nilai α yang semakin besar akan meminimalisasi efek regularisasi, memberikan model dengan kompleksitas yang lebih tinggi namun dengan potensi *overfitting* untuk *unseen data*. Selain dengan menggunakan *cross-validation* untuk *hyperparameter tuning*, teknik seperti *fractional ridge regression* [13] dengan memanfaatkan rasio L2-norm antara *regularized* dan *normal coefficients* dapat membantu menemukan nilai α yang optimal.

C. Logistic Regression (LR)

D. Decision Tree (DT)

E. Naive-Bayes (NB)

F. Linear Support Vector Machine (Linear SVM)

G. Random Forest (RF)

H. Extreme Gradient Boosting (XGBoost)

I. Categorical Boosting (CatBoost)

J. Adaptive Boosting (AdaBoost)

K. Light Gradient Boosting Machine (LGBM)

L. Ensemble (Stacking)

III. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections III-A–III-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \quad (7)$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(7)", not "Eq. (7)" or "equation (7)", except at the beginning of a sentence: "Equation (7) is . . ."

D. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Fig. 1. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

References

[1] Paper Arion
[2] Paper Arion 2
[3] Paper Arion 3
[4] Shai Shalev-Shwartz and Shai Ben-David, Understanding machine learning: From foundations to algorithms. Cambridge Etc: Cambridge University Press, 2014.
[5] F. Maymí and S. Lathrop, "AI in Cyberspace: Beyond the Hype," 2024.
[6] V. V. Putri Wibowo, Z. Rustam, and J. Pandelaki, "Classification of Brain Tumor Using K-Nearest Neighbor-Genetic Algorithm and Support Vector Machine-Genetic Algorithm Methods," in 2021 International Conference on Decision Aid Sciences and Application, Sakheer, Bahrain: IEEE, Dec. 2021, pp. 1077–1081. doi: 10.1109/DASA53625.2021.9682341.
[7] N. F. B. M. Noor, H. S. Sipail, N. Ahmad, and N. M. Noor, "Covid-19 Severity Classification Using Supervised Learning Approach," in 2021 IEEE National Biomedical Engineering Conference, Kuala Lumpur, Malaysia: IEEE, Nov. 2021, pp. 151–156. doi: 10.1109/NBEC53282.2021.9618747.
[8] A. P. Pawlovsky and H. Matsuhashi, "The use of a novel genetic algorithm in component selection for a kNN method for breast cancer prognosis," in 2017 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges, Tuxtla-Gutierrez, Mexico: IEEE, Mar. 2017, pp. 1–5. doi: 10.1109/GMEPE-PAHCE.2017.7972084.
[9] A. C. Müller and S. Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists. Beijing: O'reilly, 2017.
[10] K. Kim, "Normalized class coherence change-based k NN for classification of imbalanced data," Pattern Recognition, vol. 120, p. 108126, Dec. 2021, doi: 10.1016/j.patcog.2021.108126.
[11] K. Yuk Carrie Lin, "Optimizing variable selection and neighbourhood size in the K-nearest neighbour algorithm," Computers and Industrial Engineering, vol. 191, p. 110142, May 2024, doi: 10.1016/j.cie.2024.110142.
[12] S. Zhang, D. Cheng, Z. Deng, M. Zong, and X. Deng, "A novel k NN algorithm with data-driven k parameter computation," Pattern Recognition Letters, vol. 109, pp. 44–54, Jul. 2018, doi: 10.1016/j.patrec.2017.09.036.
[13] A. Rokem and K. Kay, "Fractional ridge regression: a fast, interpretable reparameterization of ridge regression," GigaScience, vol. 9, no. 12, p. giaa133, Nov. 2020, doi: 10.1093/gigascience/giaa133.