Reproducible and Generalizable Framework for Multi-class Hierarchical Classification Demonstrated by DNA Methylation-based Glioma Subtyping

Youdinghuan Chen[1,2,3]
[1]*National Coalition of Independent Scholars (NCIS)*
[2]*Faculty of Biological and Environmental Informatics, Wilmington University*
[3]*Dept. of Science, Mathematics and Biotechnology, University of California-Berkeley Ext.*
United States of America

Abstract

The class label of some multi-class machine-learning problems may be hierarchical. A well-known example is lower-grade glioma, which can be clinically classified first based on isocitrate dehydrogenate gene mutation and then according to chromosome 1p/19q codeletion. Hi-Class is a newer supervised framework that leverages hierarchical information within class labels for more accurate and efficient prediction. This study demonstrates the application of Hi-Class to predict the established glioma subtypes using cytosine-phosphate-guanine island methylation in two independent, real-world datasets (n=507 and n=122). As an exemplar of reproducibility, hold-out and cross-validation experiments, with or without Hi-Class, were performed. To exemplify generalizability, both Hi-Class and flat model versions were trained on one dataset and applied to another without between-cohort data processing. The strong prior knowledge and the observed classification performance make this task suitable for demonstrating and teaching specialized use cases of multi-class predictive analytics. (Keywords: Hierarchical Classification, Machine Learning, Reproducibility, Generalizability, Science Education)

1. Introduction

Machine learning (ML) prediction of multi-class labels is generally more challenging than the binary counterpart. Some multi-class scenarios involve hierarchical class labels. An example is the established subtypes of lower-grade glioma, clinically known as World Health Organization grade II or III cerebral hemisphere tumors. A glioma case is first classified as isocitrate dehydrogenase (*IDH*)-mutated if it has a mutation in one of the biologically related genes, *IDH1* and *IDH2*. *IDH*-mutated gliomas can be further classified based on whether chromosome arms 1p and 19q are lost (codeletion). This subtyping scheme is currently used as the gold standard for decision-making in the real-world clinical setting [1], [2].

Methylation is a biochemical process involving the attachment of a methyl group via covalent bonding. The fifth carbon position of DNA cytosine can be methylated. In the cytosine-phosphate-guanine (CpG) dinucleotide context of the human genome, cytosine methylation (hereafter referred to as "methylation") plays a biologically important role [3]. The most well-known functions of methylation are modulating gene expression and maintaining genome integrity [3]. Unsurprisingly, methylation alteration is an established phenomenon during cancer onset and progression [3]. In lower-grade glioma, methylation levels at genome-wide CpG islands (CGIs), which are genomic regions with a higher-than-expected density of CpG dinucleotides [3], [4], are strongly correlated with disease subtypes [1], [2], [4], [5]. Biologically, an *IDH* mutation impairs a cell's ability to maintain a normal level of methylation at CGIs across the genome, consequently yielding a distinct CGI landscape [4], [5]. This prior knowledge warrants a formal ML investigation of glioma subtypes using CGIs as predictive features.

A canonical "flat" ML classifier treats all classes equally, ignoring the hierarchical relationship within the label space. Consequently, there is a missed opportunity for more accurate and/or efficient multi-class prediction. Recently, Hi-Class was developed to capture the hierarchical information within the label space [6]. Although the superior performance and efficiency of Hi-Class were demonstrated, a unified workflow for hierarchical classification is lacking. More generally, research involving applied ML often lacks reproducibility because of the little effort to replicate well-performing models in independent studies [7].

Through a real-world case study, this report presents a recommended pipeline for multi-class hierarchical classification tasks involving high-dimensional features. The exemplary datasets used

 | DOI: 10.20533/ijcdse.2042.6364.2024.0609

were from two independent clinical populations from the real world, The Cancer Genome Atlas (referred to as the "American cohort") and the German Glioma Network (referred to as the "German cohort"). Table 1 summarizes the key attributes of the cohorts. A set of unsupervised data exploration, hold-out validation and cross-validation experimentation as well as generalizability analysis were performed. To help ease teaching and demonstration, the custom-processed CGI methylation features and curated subtypes for both cohorts were deposited at *www.kaggle.com/datasets/ydavidchen/lower-grade-glioma-cpg-islands-and-subtypes*. The R 4.3.1 and Python 3.11.5 code used for data curation, preparation, statistical and ML analysis is available on GitHub at *github.com/ydavidchen/lgg_multiclass*.

Table 1. Overview of the independent datasets with hierarchical class labels

	American n (%)	German n (%)
***IDH1* or *IDH2* gene**		
Mutated	415 (81.9)	103 (84.4)
Normal	92 (18.1)	19 (15.6)
Chromosome arms 1p and 19q		
Codeleted	167 (32.9)	36 (29.5)
Not codeleted	340 (67.1)	86 (70.5)
Subtype (class label)		
IDH wild-type (0)	92 (18.1)	19 (15.6)
IDH mutated only (1A)	248 (48.9)	67 (54.9)
IDH mutated, co-deleted (1B)	167 (32.9)	36 (29.5)

2. Data Acquisition and Processing of Two Independent Study Populations

All datasets and annotations used in this report are publicly available without access control. The genome-wide methylation data (input features) and glioma subtypes (classification labels) of the American cohort were curated from publicly available repositories SynapseTCGALive (accession syn1910181) and cBioPortal (dataset LGG-Pancancer Atlas), respectively [8], [9]. The methylation data and subtypes of the German cohort are publicly available in the Gene Expression Omnibus database (accession GSE129477) [2], [10]. The methylation platform annotation necessary for preparing the ML feature space is publicly available from the Illumina product support webpage.

The class labels of the cohorts are hierarchical. That is, a glioma sample can be classified as *IDH*-normal, *IDH*-mutated, or *IDH*-mutated with chromosome 1q/19q codeletion. The "outer class" is the binary status of *IDH* mutation assigned to all samples. The "inner class" is the binary codeletion status available only to the *IDH*-mutated subset. The feature space consists of the average methylation profiles of gene-associated autosomal CGIs. To prepare the feature space, the data matrix at the probe level was first filtered to contain only the autosomal measurements. A typical methylation dataset follows a bimodal distribution ranging from 0 (no methylated alleles) to 1 (all alleles methylated). To improve downstream ML, the data was logit-transformed to approximate a Gaussian distribution [11]. In addition, only the CGIs with a minimum of 800 bases (median=825, mean=1,015) measured by at least six probes (median=5, mean=5.59) were kept. Further, only CGIs located within annotated gene promoters (designated "TSS200" and "TSS1500") were kept due to their putative biological relevance [3]. The processed feature matrix was then joined with the manufacturer's annotation on the probe identifiers, grouped by unique CGIs, and mean-aggregated. The final data matrix contained logit-transformed methylation values of 3,017 gene promoter-associated CGIs across *n* samples ($n_{American} = 507$, $n_{German} = 122$) with known subtypes as labels for classification. To avoid unfairly biasing the model performance via data leakage [7], the dataset cohorts were independently prepared. There was no between-dataset cross-normalization of the feature space.

3. Unsupervised Data Exploration to Confirm the Feasibility of Supervised ML

The relationship between the feature space and the classification labels is the prerequisite for building robust ML models. To verify the positive relationship between the CGI features and the disease subtypes, two unsupervised analyses were conducted.

3.1 Principal Component Analysis (PCA)

The first unsupervised approach was PCA, applied to the entire feature space (i.e. all CGIs) with the R function *prcomp*. Figure 1 visualizes the topmost principal components carrying the greatest and unrelated variance of the CGI feature space [12]. Here, PCA confirmed a clear distinction by subtype in both cohorts offering a basis for ML.

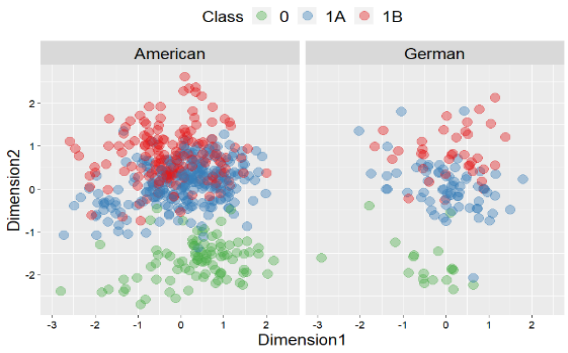

Figure 1. Feature space with standard-scaled PCA Class

Figure 1, exploration of the feature space with standard-scaled PCA. Class labels *0*, *1A*, and *1B* represent *IDH*-normal, *IDH*-mutated without and with chromosome 1p/19q codeletion.

3.2. Unsupervised Hierarchical Clustering (UHC)

To further support the feasibility of supervised ML, the most variable 750 (24.9%) CGIs, constrained within values ±6.0, were subject to UHC with base R's *hclust* with Euclidean distance and Ward D linkage, and then visualized with *pheatmap* (R package v.1.0.12) [13], [14]. As shown in Figure 2, the UHC patterns of the two independent cohorts were strikingly similar despite the sample size difference.

To quantify the positive association between the feature and label space, each sample was assigned to a UHC cluster by splitting the dendrogram tree at the second node. Both cohorts had a distinct, smaller *Cluster 0* and a larger *Cluster 1* with *Sub-clusters 1A* and *1B*. Fisher's exact tests applied to the 3x3 contingency tables revealed a significant overlap between the UHC clusters and the classes (P_{American} = 8.1e-136 and P_{German} = 1.6e-30). Taken together, the consistent PCA and UHC results helped verify the potential of supervised ML with robustness.

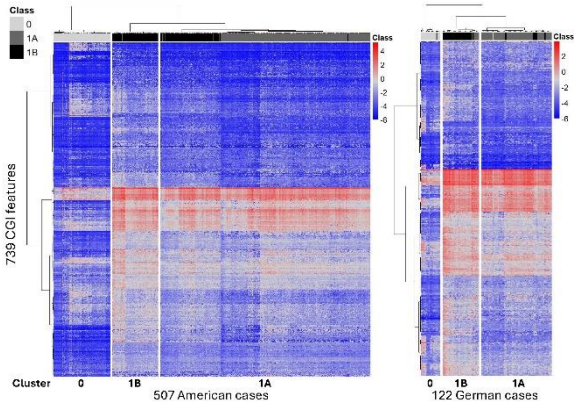

Figure 2. UHC of the Top Quarter most Variable CGI Features

Figure 2, shows the UHC of the top quarter most variable CGI features. Rows and columns are CGI and glioma tumors, respectively. Class labels *0*, *1A*, and *1B* represent *IDH*-normal, *IDH*-mutated without and with chromosome 1p/19q codeletion.

4. Experimentation of Predicting Hierarchical Class Labels in Independent Study Populations Aiming for Reproducibility

Since the unsupervised analyses confirmed the previously known, strong relationship between CGIs and glioma subtypes [1], [2], [4], [5], the next step would be leveraging supervised ML for CGI-based subtype prediction. All multi-class ML models and evaluation metrics were implemented in Python *scikit-learn* [15]. To quantitatively determine the predictive capability of multi-class models, an *in silico* experiment with a 2x2 design was performed: two evaluation modes (hold-out and 5-fold cross-validation) with or without the Hi-Class wrapper. The logistic regression classifier with LASSO penalty commonly used in high-dimensional genomics research [16] was chosen for both Hi-Class and flat classifiers. Gaussian Support Vector Machine (SVM) classifiers were built in parallel. Each modeling attempt was evaluated with metrics including accuracy and balanced accuracy (i.e. mean accuracy per class) as well as class-weighted precision, recall, and F1 score (i.e. harmonic mean of precision and recall) [15]. Tables 2 and 3 summarize the experimental results.

4.1. Hold-out Validation

To assess whether the Hi-Class and flat multi-class ML could accurately predict glioma subtypes from CGI profiles on unseen observations, a hold-out experiment was performed. Each dataset was split into a training and a test set at an 80:20 ratio. The ML models were built on the training and evaluated on the test set. The experiment was conducted in both cohorts, with or without the Hi-Class wrapper. Table 2 shows near-perfect performance across all modeling attempts and both cohorts.

Table 2. Hold-out experiment results

	American		German	
	Hi-Class	Flat	Hi-Class	Flat
(A) LASSO				
Accuracy	99.0%	99.0%	100%	100%
Balanced Accuracy, Precision, Recall, F1 Score	99.2%	99.2%	100%	100%
(B) Gaussian SVM				
Accuracy	99.0%	99.0%	100%	100%
Balanced Accuracy, Precision, Recall, F1 Score	99.2%	99.2%	100%	100%

4.2 Cross Validation

To validate the observed model performance and assess classifier stability, a *K*-fold cross-validation was performed. Each cohort was split into *K*=5 folds. *K*-1=4 folds were used for model development and the remainder was used for evaluation. The procedure was repeated *K* times to ensure every fold would be

 | DOI: 10.20533/ijcdse.2042.6364.2024.0609

used once. As shown in Table 3, the model performance in the cross-validation experiment was strong and consistent overall. Hi-Class slightly improved the performance. The average metrics across *K* folds were strong and similar to those of the hold-out experiment described earlier.

Table 3. Cross-validation experiment results

	American		German	
	Hi-Class	Flat	Hi-Class	Flat
(A) LASSO				
Accuracy	**98.4%**	98.2%	**94.3%**	93.4%
Bal. Accuracy	**98.8%**	98.6%	**93.2%**	92.3%
Precision	**98.6%**	98.3%	**95.1%**	94.7%
Recall	**98.8%**	98.6%	**93.2%**	92.3%
F1 Score	**98.6%**	98.4%	**93.5%**	92.7%
(B) Gaussian SVM				
Accuracy	98.8%	98.8%	**93.5%**	91.8%
Bal. Accuracy	99.2%	99.2%	**92.7%**	90.2%
Precision	98.8%	98.8%	**94.3%**	93.8%
Recall	99.2%	99.2%	**92.7%**	90.2%
F1 Score	99.0%	99.0%	**92.8%**	90.9%

5. Generalizability Analysis via Direct Application of Models without Between-Cohort Data Processing

To achieve a real-world impact, ML models need to display generalizability and external validity by well performing in unseen data never exposed to the training process [17]. Therefore, the ML models trained on the American cohort (n=507) were applied directly to the German cohort (n=122), and *vice versa*. To prevent data leakage [7], the feature space of each cohort was pre-processed independently without contaminating one another.

Table 4. Generalizability analysis results.

	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
(A) Trained on American, applied to German					
All models and versions	95.9%	95.0%	96.3%	95.0%	95.6%
(B) Trained on German, applied to American					
LASSO, Hi-Class	**95.7%**	**95.9%**	96.4%	95.9%	96.1%
LASSO, flat	95.9%	96.3%	96.4%	96.3%	96.3%
Gaussian SVM, Hi-Class	**97.0%**	**97.5%**	**97.3%**	**97.5%**	**97.4%**
Gaussian SVM, flat	94.5%	94.7%	95.8%	94.7%	95.1%

Table 4 shows that either scenario decreased little in performance compared to the previous experiments. When trained on the American cohort with a larger sample size, model performance was stronger and more consistent, as expected. Surprisingly, when trained on the German cohort with a nearly fourfold reduction in sample size, there was little loss in performance metrics.

Notably, the classifiers chosen were already high-scoring in general, and thus the room for improvement was expected to be limited. Nevertheless, Hi-Class led to improved classification when Gaussian SVM was trained on the less statistically powered cohort.

6. Discussion

This report presents a real-world case study of multi-class hierarchical classification with Hi-Class implementation [6]. The chosen machine learning (ML) task involving the relationship between cytosine-phosphate-guanine islands (CGI) methylation and lower-grade glioma subtypes has been established [1], [2]. The classification label of this report, glioma subtypes, has a hierarchical nature and a known correlation with the feature space. Thus, the ML task at hand was suitable for demonstrating and teaching Hi-Class and related approaches. This report subsequently presents a series of recommended steps to address the task, including unsupervised data exploration, supervised hold-out and cross-validation experimentation as well as a generalizability analysis.

This report highlights two key principles that are often lacking in applied ML domains [7]. First, ML models should be trained with reproducibility. Here, the hold-out and cross-validation experiments were replicated in two independent datasets and showed similarly strong performance. Second, effort should be made to improve the generalizability of developed models. Here, the classifiers, with or without Hi-Class, trained on one cohort were applied to another directly. There was little loss in performance despite the differences in feature-space distributions and cohort characteristics which were left as-is. Notably, by independently processing the cohorts, the present study effectively prevented data leakage, the unwarranted and often inadvertent inclusion of external information for classification with the potential of unfairly biasing model performance [7].

This study has limitations. Despite the strong ML performance in general, the present study was underpowered compared to ML problems in other scientific domains, as is the case for biomedical data in general. Future ML works for pedagogical and demonstrative purposes should exemplify the principles shown in this work in larger datasets. This may help enable the use of deep learning-based frameworks known to exceed the performance of conventional ML and detect more nuanced feature-label relationships [18].

One area needing further investigation is the Hi-Class feature importance. In the context of the domain investigated, interpreting the relevance of critical CGI features will likely have diagnostic and prognostic utility. Finally, ongoing efforts should evaluate hierarchical and flat classification with more

 | DOI: 10.20533/ijcdse.2042.6364.2024.0609

challenging ML problems, such as those with noisy feature space less correlated to the classification labels. It will be interesting to demonstrate Hi-Class applications in future case studies where the hierarchical framework is considerably superior to canonical classification.

7. References

[1] Cancer Genome Atlas Research Network et al., 'Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas.', N Engl J Med, vol. 372, no. 26, pp. 2481–98, Jun. 2015, DOI: 10.1056/NEJMoa1402121.

[2] H. Binder et al., 'DNA methylation, transcriptome and genetic copy number signatures of diffuse cerebral WHO grade II/III gliomas resolve cancer heterogeneity and development', Acta Neuropathol Commun, vol. 7, no. 1, p. 59, Dec. 2019, DOI: 10.1186/s40478-019-0704-8.

[3] P. A. Jones, 'Functions of DNA methylation: islands, start sites, gene bodies and beyond.', Nat Rev Genet, vol. 13, no. 7, pp. 484–92, 2012, DOI: 10.1038/nrg3230.

[4] J. Yates and V. Boeva, 'Deciphering the etiology and role in oncogenic transformation of the CpG island methylator phenotype: a pan-cancer analysis', Brief Bioinform, vol. 23, no. 2, Mar. 2022, DOI: 10.1093/bib/bbab610.

[5] S. Turcan et al., 'IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype.', Nature, vol. 483, no. 7390, pp. 479–83, Feb. 2012, DOI: 10.1038/nature10866.

[6] F. M. Miranda, N. Köhnecke, and B. Y. Renard, 'HiClass: a Python Library for Local Hierarchical Classification Compatible with Scikit-learn', Journal of Machine Learning Research, vol. 24, no. 29, pp. 1–17, 2023.

[7] S. Kapoor and A. Narayanan, 'Leakage and the reproducibility crisis in machine-learning-based science', Patterns, vol. 4, no. 9, p. 100804, Sep. 2023, DOI: 10.1016/j.patter.2023.100804.

[8] J. Gao et al., 'Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal', Sci Signal, vol. 6, no. 269, pp. pl1–pl1, 2013, DOI: 10.1126/scisignal.2004088.

[9] J. N. Weinstein et al., 'The Cancer Genome Atlas Pan-Cancer analysis project', Nat Genet, vol. 45, no. 10, pp. 1113–1120, Oct. 2013, DOI: 10.1038/ng.2764.

[10] T. Barrett et al., 'NCBI GEO: archive for functional genomics data sets—update', Nucleic Acids Res, vol. 41, no. D1, pp. D991–D995, Nov. 2012, DOI: 10.1093/nar/gks1193.

[11] P. Du et al., 'Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis', BMC Bioinformatics, vol. 11, no. 1, p. 587, Dec. 2010, DOI: 10.1186/1471-2105-11-587.

[12] J. Shlens, 'A Tutorial on Principal Component Analysis', ArXiv, vol. abs/1404.1100, 2014, DOI: 10.48550/arXiv.1404.1100.

[13] Y. Chen, 'Transcriptomic Profiling of Subcutaneous Adipose Tissue in Relation to Bariatric Surgery: A Retrospective, Pooled Re-analysis.', J Obes Metab Syndr, vol. 32, no. 1, pp. 98–102, Mar. 2023, DOI: 10.7570/jomes22065.

[14] Y. Chen, 'Pooled microarray expression analysis of failing left ventricles reveals extensive cellular-level dysregulation independent of age and sex', Journal of Molecular and Cellular Cardiology Plus, vol. 7, p. 100060, 2024, DOI: 10.1016/j.jmccpl.2023.100060.

[15] F. Pedregosa et al., 'Scikit-learn: Machine Learning in Python', Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825–2830, 2012.

[16] R. Tibshirani, 'Regression Selection and Shrinkage via the Lasso', Journal of the Royal Statistical Society B, vol. 58, no. 1. pp. 267–288, 1996. DOI: 10.2307/2346178.

[17] Y. Chen et al., 'Molecular and epigenetic profiles of BRCA1-like hormone-receptor-positive breast tumors identified with development and application of a copy-number-based classifier.', Breast Cancer Res, vol. 21, no. 1, p. 14, 2019, DOI: 10.1186/s13058-018-1090-z.

[18] M. Wainberg, D. Merico, A. Delong, and B. J. Frey, 'Deep learning in biomedicine', Nat Biotechnol, vol. 36, no. 9, pp. 829–838, Oct. 2018, DOI: 10.1038/nbt.4233.

