Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 1, February 2026, pp. 568~584
ISSN: 2302-9285, DOI: 10.11591/eei.v15i1.9087

Improving genomic classification via Pearson-based SNP selection: a comparison of k-NN, SVM, and random forest

Prihanto Ngesti Basuki[1], Sri Yulianto Joko Prasetyo[1], Adi Setiawan[2]
[1]Faculty of Information Technology, Satya Wacana Christian University, Salatiga, Indonesia
[2]Faculty of Science and Mathematics, Satya Wacana Christian University, Salatiga, Indonesia

Article Info

Article history:

Received Jul 31, 2024
Revised Oct 29, 2025
Accepted Dec 6, 2025

Keywords:

Ancestry inference
High-dimensional data (small-n, high-p)
Leakage-free evaluation
Monte Carlo cross-validation
Point-biserial correlation
Principal component analysis
Receiver operating-characteristic area under the curve

ABSTRACT

Accurate genomic classification is vital for precision health and population studies, yet high-dimensional single-nucleotide polymorphism (SNP) data ($p \gg n$) amplify noise, redundancy, and overfitting. This study evaluates a simple, model-independent Pearson-based selection that ranks SNPs by feature–label correlation, and assesses k-nearest neighbors (k-NN), linear support vector machine (SVM), and random forest (RF) under leakage-free stratified Monte Carlo cross-validation (MCCV). Performance increases monotonically with $|r|$: the strongest tiers reach ≈99–100% accuracy; SVM leads in mid tiers (RF second), while k-NN is competitive mainly at the extremes. A matched-dimensionality PCA-120 baseline (TRAIN-only) attains parity for SVM/RF and trails slightly for k-NN at the 10% test size. With 120-SNP panels, prediction medians are ≈0.30 ms (SVM), 1.81–1.83 ms (k-NN), and 34–35 ms (RF), supporting CPU-only deployment. A consensus panel combining correlation evidence with principal component analysis (PCA) selection frequency yields interpretable Top-20/Top-120 subsets and $|r|$-based operating thresholds. Overall, Pearson-based selection provides a transparent, reproducible baseline for small-sample SNP classification, offering accuracy competitive with PCA at lower computational complexity and straightforward extensions to broader cohorts and multi-omics integration.

Corresponding Author:

Prihanto Ngesti Basuki
Faculty of Information Technology, Satya Wacana Christian University
St. Dr. O. Notohamidjojo No. 1 - 10, Blotongan, Kec. Sidorejo, Salatiga, Central Java, Indonesia
Email: ngesti@uksw.edu

1. INTRODUCTION

Approximately 99.9% of the human genome is identical across individuals, and the remaining ≈0.1% consists largely of single-nucleotide polymorphisms (SNPs) occurring about once every ≈1,000 bases [1]. These variants contribute to evolution, pharmacogenomic response, and risk for complex diseases (e.g., obesity, diabetes, hypertension, and cancer) [2]–[6]. They are widely used in biomedical and population studies, forensic inference, and phenotype prediction [7]–[10].

Machine learning classifiers are commonly applied to SNP data. k-nearest neighbors (k-NN) assigns labels by local majority vote [11], [12]; support vector machine (SVM) constructs maximum-margin hyperplanes [13]; and random forest (RF) aggregates decision trees for improved stability [14]. These methods have shown effectiveness across diverse application areas, including disease prediction, cybercrime detection, GPS data analytics, exam classification, and genetic studies [15]–[22].

Genomic datasets frequently exhibit a high-dimensional $p \gg n$ regime—far more features (p) than samples (n)—which elevates overfitting risk and motivates feature selection to reduce dimensionality and

preserve interpretability [23]. Principal component analysis (PCA) is a popular baseline for dimensionality reduction and population-structure control [24]–[27]. Recent evaluations also highlight caveats: adjusting for principal components can induce collider bias in GWAS models [24], and population structure inferred by PCA can diverge from admixture-model estimates [25]. However, components may not align with predictive relevance and are difficult to interpret at the locus level [28]–[31]. In contrast, Pearson-based selection—ranking SNPs by the Pearson/point-biserial correlation with the label—offers a simple, fast, and transparent filter [23], [32] that preserves the original SNP representation, making the feature–label relationship explicit and scaling well to high-dimensional genomic data [33], [34]. Prior studies indicate that Pearson-based SNP selection can improve classification in small-sample settings [35].

The present study evaluates Pearson-based filtering for SNP classification on HapMap Phase II (9,305 SNPs; CEU vs YRI) under stratified, leakage-free Monte Carlo cross-validation (MCCV) using k-NN, linear SVM, and RF; MCCV is employed to obtain robust performance estimates via repeated randomization of test partitions [36]. The contribution is fourfold: i) |r|-tiering as a difficulty index, ii) leakage-free MCCV with $p \approx n$ block sizing (≤120 features) to curb overfitting, iii) a matched-dimensionality PCA-120 baseline constructed from TRAIN-only PCA, scoring SNPs by loading magnitudes weighted by each component's explained variance ratio (EVR), and iv) a consensus panel (Top-20 in the main text and Top-120/240 available). Together, these elements yield an interpretable, reproducible template for small-sample SNP classification while retaining locus-level interpretability.

From an informatics and signal-processing perspective, SNP classification is a supervised pattern-recognition problem in high-dimensional noise. Pearson-based selection serves as a lightweight, linear pre-processing step that improves the signal-to-noise ratio and preserves locus-level interpretability, with computation that scales linearly with the number of features. At matched dimensionality (120 features), inference on standard CPUs operates at millisecond to sub-second scale for k-NN, linear SVM, and RF, enabling deployment on edge or clinical workstations (see subsection 4.5). This framing highlights an accuracy–efficiency balance suitable for real-time or resource-constrained settings.

2. RELATED WORK

Filter, wrapper, and embedded approaches are widely used for SNP selection. Single-marker filters such as Pearson, point-biserial, chi-square (χ^2), and mutual information are fast and model-independent, but ignore LD and interactions [37], [38]. Recursive feature elimination (RFE) iteratively removes low-contribution variables relative to a target classifier (e.g., SVM-RFE or RF-RFE), improving accuracy but requiring repeated model fitting [39], [40]. Embedded methods (L1-regularized logistic/linear SVM) select features via sparsity but are model-dependent [41], [42]. GWAS-style per-SNP tests with multiple-testing control (e.g., Bonferroni, false discovery rate (FDR)) offer interpretable thresholds yet are not always optimal for classification [22]. Projection methods (PCA/PLS) aid structure control but reduce interpretability; variance captured does not guarantee predictive relevance [24]–[31].

In this context, the evaluation centers on a leakage-free Pearson filter contrasted with a PCA-120 baseline to isolate the value of label-aware selection versus unsupervised projection. LD-aware redundancy control and nested resampling are acknowledged as extensions beyond the present scope.

3. METHOD

3.1. Research workflow

The end-to-end workflow is summarized in Figures 1 and 2. Figure 1 summarizes QC (MAF≥0.05; HWE; missingness ≤10%), dosage encoding (0/1/2), Pearson-based selection with TRAIN-fold minor-allele mapping, and TRAIN-only median imputation. Figure 2 summarizes leakage-free MCCV (R=1000; test sizes 10/25/40), the evaluated models (k-NN, linear SVM, RF100/125), key metrics (incl. ROC–AUC), and statistical validation. The procedure comprises: i) data quality control (QC) and genotype encoding; ii) Pearson correlation calculation for each SNP against the binary label; iii) ranking SNPs by correlation and separating them into positive and negative groups; iv) partitioning each group into sub-datasets of 120 SNPs (correlation blocks); v) classification with k-NN, linear SVM, RF100, and RF125 under MCCV (R=1000) at test sizes 10%, 25%, and 40% [43]; vi) evaluation using accuracy, precision, recall, F1, receiver operating characteristic (ROC), area under the curve (AUC) and normalized confusion matrices (CM); vii) statistical validation: Shapiro–Wilk; if assumptions held, one-way ANOVA with Tukey HSD; otherwise Kruskal–Wallis with Bonferroni-adjusted pairwise Mann–Whitney U [44]; and viii) identification of SNP loci associated with peak performance within specific correlation ranges.

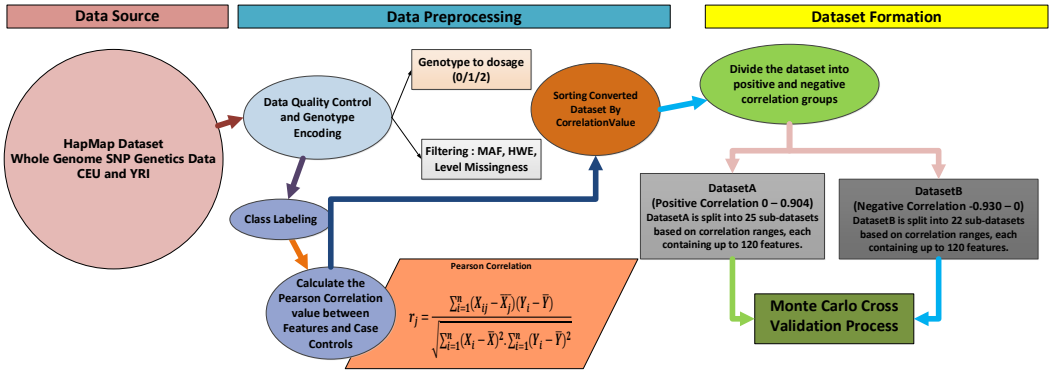


Figure 1. Workflow for SNP selection using feature–target correlation (Part 1 of 2)

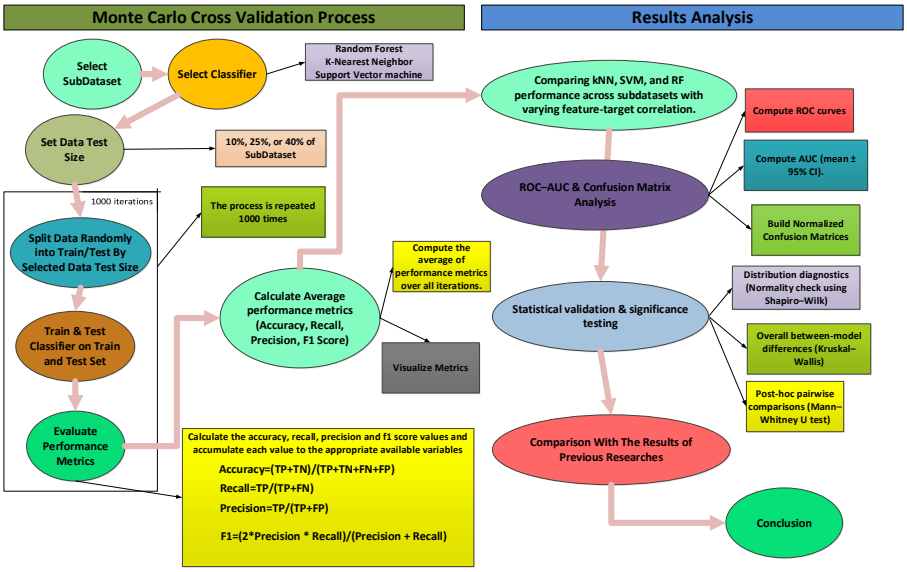


Figure 2. Workflow for model training, evaluation, and statistical validation (Part 2 of 2)

To focus the analysis, a model-independent filter is adopted that ranks SNPs by Pearson feature–target correlation; redundancy-aware filters (e.g., mRMR) are not considered. A PCA baseline with 120 components is included to compare projection-based reduction against correlation-ranked subsets.

3.2. Research data

The dataset was obtained from HapMap Phase II (2007), comprising 120 individuals—60 CEU (Utah residents of European ancestry) and 60 YRI (Yoruba in Ibadan, Nigeria)—and 9,305 SNPs [45]. Data handling and association analyses were performed in R with SNPassoc [46]. Genotypes are encoded from the four nucleotides adenine (A), cytosine (C), guanine (G), and thymine (T). Some loci contain missing genotype calls due to weak genotyping signals, platform limitations, DNA quality issues, allelic dropout, or calling errors [47]. Missingness can be informative, and genotype imputation may introduce bias when mechanisms are non-random or coverage is limited [48], [49]. A small subset of the data is shown in Table 1. Row headers are dbSNP rsIDs ("rs" = Reference SNP cluster ID, NCBI dbSNP) [50]; column headers are individual sample IDs (e.g., NA06985), each uniquely identifying a CEU or YRI subject.

Table 1. Subset of SNP genotypes data from CEU and YRI populations in the HapMap Phase II dataset

Samples	NA06985	NA06993	NA06994	NA19116	NA19119	NA19127
Groups	CEU	CEU	CEU	YRI	YRI	YRI
rs11260616	AA	AT	AA	AA	AT	AA
rs6659552	GG	CG	CG	GG	GG	GG
rs6688969	CC	CT	CT	CT	CT	CC
rs10753357	AC	AA	AA	AC	CC	AC

3.3. Data quality control and genotype encoding

3.3.1. Data quality control

The raw SNP dataset was subjected to QC filtering to ensure data reliability. Variants with minor allele frequency (MAF)<0.05, Hardy–Weinberg Equilibrium (HWE) $p < 1 \times 10^{-6}$, or missingness >10% were excluded. These thresholds mitigate known risks—low MAF shrinks genotype variance ($Var[G]=2p(1-p)$) yielding unstable correlations and fragile decision boundaries [51]; HWE deviations may indicate genotyping error or substructure that biases correlation ranks and PCA loadings. Recent reassessments also show that HWE filtering can alter inferred population structure [52], [53]; and high missingness reduces effective sample size, suggests non-random loss, forces heavier imputation, and inflates uncertainty [54], [55]. When missingness is non-random or coverage is limited, genotype imputation can introduce systematic bias in downstream association and classification [48], [49]. Applying QC prior to Pearson correlation ranking and PCA follows common GWAS practice, helping ensure downstream results reflect biological signal rather than data-quality artifacts. After QC, 5,647 SNPs remained from the initial set; variants failing criteria—along with zero-correlation loci—were removed so that only high-quality markers proceeded to feature selection and classification.

3.3.2. Imputation after quality control

With per-SNP missingness ≤10%, missing values in dosage-coded genotypes (0/1/2) were filled with the per-SNP median as a simple single-imputation step to avoid heavy model-based imputation and information leakage across MCCV folds (imputation parameters are computed on training folds and applied to the held-out test data)—an accepted practice that also stabilizes PCA inputs [56]; notably, PLINK 2 mean-imputing by design for PCA [57]. LD/haplotype-based imputation was not used because the goal was to fill sparse missing calls rather than infer untyped variants [58].

3.3.3. Encoding and correlation

Each SNP was encoded into allele dosages (0/1/2) relative to the minor allele; e.g., if "A" is minor, then CC=0, AC=1, and AA=2. Missing genotypes were set to NaN, imputed with the training-set median per SNP, and the same mapping was applied to the test split to avoid information leakage. The encoding and imputation step is summarized in Figure 3.

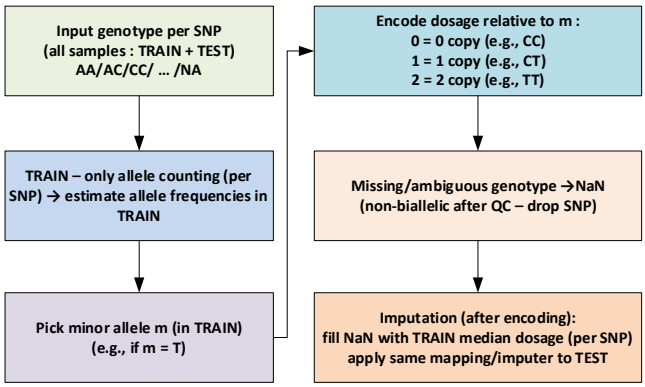


Figure 1. Encoding genotypes as allele dosages (0/1/2) with median imputation fitted on the training set

To quantify linear association with the population (CEU=0, YRI=1), Pearson's correlation was calculated per SNP: for SNP j, let x_{ij} denote the encoded dosage of sample i, $\bar{x_j}$ its mean, y_i the binary label, and $\bar{y}$ its mean; the coefficient r_j is given in (1):

$$r_j = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 . \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

3.3.4. Ordering and subset construction

SNPs were ranked by descending Pearson correlation (from strongest positive to strongest negative) to define Pearson-based subsets. SNPs with $r = 0$ were excluded—only 12 of 5,647 (≈0.21%), so the impact was negligible. The ordered list appears in Table 2.

Table 2. SNPs sorted in descending Pearson correlation (r) with the population label (positive to negative)

Features	X_1	X_2	...	X_{3009}	...	X_{5646}	X_{5647}
Pearson correlation (r)	0.9039	0.9038	...	0	...	-0.904	-0.929
Reference number	rs9909962	rs2370893	...	rs12928849	...	rs6670842	rs10868791

3.4. Research stages

Based on Table 2, data were split into DatasetA ($r > 0$; X_1-X_{2997}, range (0.000–0.904]) and DatasetB ($r < 0$; X_{3010}-X_{5647}, range [-0.929–0.000)). SNPs with $r = 0$ (X_{2998}-X_{3009}) were excluded, since they provide no linear association with the label. Correlation ranges are left-inclusive and right-exclusive, except for the last interval which is closed on both ends.

To balance experiments, each group was partitioned into ≤120-SNP sub-datasets so that $p \approx n$=120, reducing overfitting risk and enabling fair comparison across correlation ranges. This produced 25 sub-datasets for DatasetA (Table 3) and 22 for DatasetB (Table 4). Correlation intervals are left-inclusive, right-exclusive (the last interval is closed). Sub-datasets are referenced by block IDs (e.g., A1, B1) that encode fixed correlation ranges (positive for A-blocks, negative for B-blocks); *strong*, *moderate*, and *weak* denote upper-, mid-, and lower-correlation tiers.

Table 3. Splitting DatasetA (positive correlations) into sub-datasets of up to 120 features

Blocks	Correlation range	Blocks	Correlation range	Blocks	Correlation range	Blocks	Correlation range	Blocks	Correlation range
A1	[0.673-0.904]	A6	[0.488-0.522)	A11	[0.360-0.377)	A16	[0.255-0.281)	A21	[0.114-0.141)
A2	[0.601-0.673)	A7	[0.463-0.488)	A12	[0.342-0.360)	A17	[0.224-0.255)	A22	[0.088-0.113)
A3	[0.558-0.601)	A8	[0.437-0.463)	A13	[0.324-0.342)	A18	[0.199-0.224)	A23	[0.058-0.088)
A4	[0.522-0.557)	A9	[0.418-0.437)	A14	[0.306-0.324)	A19	[0.171-0.198)	A24	[0.029-0.058)
A5	[0.488-0.522)	A10	[0.397-0.418)	A15	[0.282-0.306)	A20	[0.141-0.171)	A25	(0.000-0.029)

Table 4. Splitting DatasetB (negative correlations) into sub-datasets of up to 120 features

Blocks	Correlation range	Blocks	Correlation range	Blocks	Correlation range	Blocks	Correlation range	Blocks	Correlation range
B1	[-0.029-0.000)	B6	[-0.168--0.140)	B11	[-0.318--0.286)	B16	[-0.444--0.420)	B21	[-0.678--0.603)
B2	[-0.058--0.029)	B7	[-0.196--0.168)	B12	[-0.344--0.318)	B17	[-0.472--0.444)	B22	[-0.929--0.678)
B3	[-0.083--0.058)	B8	[-0.223--0.196)	B13	[-0.367--0.345)	B18	[-0.507--0.473)		
B4	[-0.111--0.083)	B9	[-0.254--0.223)	B14	[-0.392--0.367)	B19	[-0.549--0.507)		
B5	[-0.140--0.112)	B10	[-0.285--0.254)	B15	[-0.420--0.393)	B20	[-0.603--0.549)		

3.5. Feature selection (Pearson, principal component analysis) and consensus panel

Pearson correlation was computed between each SNP (dosage 0/1/2) and the binary label; SNPs were ranked by descending $|r|$ and grouped into positive (DatasetA) and negative (DatasetB) blocks of ≤120 SNPs. As an unsupervised baseline, PCA was applied to the post-QC genotype matrix. Within each MCCV repetition, genotype matrices were standardized and PCA was fitted on the training folds only to prevent information leakage. The PCA baseline derives an unsupervised SNP ranking via EVR-weighted loading magnitudes and retains the top-120 loci (PCA-120)—rather than using components as features. An EVR-weighted SNP score is computed as $s_j = \sum_c \mathrm{EVR}_c \cdot |\,\mathrm{loading}_{j,c}\,|$ using TRAIN-only PCA.

A consensus SNP panel was then defined using a block-agnostic rule: variants received a combined score equal to the average of; i) the percentile of $|r|$ and ii) the percentile of PCA selection frequency across MCCV resamples. Presentation and availability of the Top-20/50/120/240 panels and the full ranked table are described in section 4.6.

3.6. Computational complexity (filter vs principal component analysis vs wrappers)

Let n denote the number of samples and p the number of SNPs; $k = 120$ for the PCA baseline. Pearson-based selection computes the point-biserial correlation for all p SNPs in $O(n \cdot p)$ time and ranks them in $O(p \log p)$, with memory scaling linearly in p. The transform is fitted on TRAIN within each MCCV resample and then applied unchanged to TEST to avoid leakage. PCA (train-only) to k components via truncated/randomized SVD scales approximately as $O(n \cdot p \cdot k + p \cdot k^2)$ (exact SVD: $O(\min\{n \cdot p^2, p \cdot n^2\})$); projection of TEST uses the TRAIN-fitted components and requires $O(p \cdot k)$ memory. Wrapper methods (e.g., recursive/forward selection) incur repeated model fitting with cost $\approx O(R \times \mathrm{CV} \times C_{\mathrm{train}})$, where R is the number of elimination/forward rounds, CV the inner folds/repeats, and

C_{train} the base-learner training cost (for reference, linear SVM $\sim O(n \cdot p)$ per pass; RF $\sim O(T \cdot n \log n)$ for T trees). In the $p \gg n$ regime typical of SNP data, Pearson-based selection is markedly lighter than wrappers and generally lighter than PCA at $k = 120$, while preserving locus-level interpretability; empirical runtimes (reported in subsection 4.5) are consistent with these order-of-growth expectations.

3.7. Classification and evaluation

3.7.1. Classification

Four classifiers were applied—RF100, RF125, linear SVM, and k-NN (Euclidean). Classification used stratified MCCV with test sizes of 10%, 25%, and 40%, repeated R=1,000 times per block and test size. In each repetition, data were split into training and test sets; encoding and per-SNP median imputation were fit on the training split and applied to the test split to prevent leakage. For k-NN, *k* was chosen by inner stratified 5-fold CV on the training set to maximize accuracy, whereas SVM and RF used fixed hyperparameters. The evaluation loop, including the inner-CV scheme for k-NN, is summarized in Algorithm 1.

Algorithm 1. Stratified MCCV evaluation (SVM/RF fixed; k-NN via inner-CV)

```
Inputs:
 Blocks S in {A1..A25, B1..B22}
 Test sizes T = {0.10, 0.25, 0.40}
 Repeats R = 1000
 Classifiers C_fixed = {linear SVM, RF100, RF125}
 Label y in {0,1}
for each block S:
 for each t in T:
 repeat r = 1..R:
 Stratified split with test size t:
 S -> (X_train, y_train), (X_test, y_test)
 Fit encoding & per-SNP median imputation on (X_train); apply to (X_test)
 # Fixed-parameter models
 for c in C_fixed:
 Train c on (X_train, y_train); predict on X_test; record metrics
 # k-NN with inner CV on the training set
 Choose k* = argmax_k Accuracy via stratified V-fold CV (V = 5) on (X_train)
 Train k-NN(k*) on full (X_train, y_train); predict on X_test; record metrics
Aggregate: compute mean +/- SD and median [IQR] over R; retain per-model distributions for
statistical tests.
```

3.7.2. Metrics evaluation

The following metrics were recorded on the held-out test sets: accuracy, precision, recall, F1-score, ROC–AUC, normalized CM, and execution time. Results were summarized as mean ± SD across repetitions and median [IQR]. Visualizations comprised correlation-tier metrics, cross-block accuracy summaries, ROC curves, and normalized CM; AUC with confidence intervals accompanied the plots. Between-classifier comparisons used the inferential procedures described in subsection 3.7.

3.8. Statistical validation

Statistical validation was conducted to examine the significance of performance differences among classifiers. The analysis was applied separately for each sub-dataset and test size, using Accuracy as the primary metric. Other metrics, including precision, recall, and F1-score, were also summarized descriptively to provide a broader view of classifier performance.

Normality of the metric distributions was first assessed using the Shapiro–Wilk test at a significance level of α=0.05. If all groups satisfied normality and variance homogeneity (Levene's test, α=0.05), a one-way ANOVA was performed, followed by Tukey's HSD for post-hoc pairwise comparisons. When assumptions were violated, a Kruskal–Wallis test was used as the non-parametric alternative. If significant, post-hoc pairwise testing was conducted using Mann–Whitney U tests (Bonferroni-adjusted) for multiple comparisons. Descriptive statistics (mean±SD, median [IQR]) accompanied the inferential results to aid interpretation.

4. RESULTS AND DISCUSSION

4.1. Performance on DatasetA (positively correlated SNPs)

This section reports DatasetA (positive-correlation) results by correlation tier and test size. Representative results at 25% test size are presented in Table 5, which retains the full set of nine correlation blocks (A1, A4, A7, A10, A13, A16, A19, A22, and A25) spanning strong→weak correlations.

Table 5. DatasetA: average accuracy (%) of RF, k-NN, and SVM with 25% test data

Pearson corr. range	A1	A4	A7	A10	A13	A16	A19	A22	A25
k-NN	100	98.78	84.33	67.6	62.1	68.07	64.05	51.48	36.56
SVM	100	100	100	99.96	98.49	97.32	84.22	56.91	17.38
RF100	100	100	100	99.39	97.58	92.35	77.7	51.95	24.64
RF125	100	100	100	99.53	97.75	93.16	78.7	52.04	23.33

4.1.1. Accuracy and metric trends

Accuracy increases monotonically with correlation strength: in the strongest tiers (A1–A7) all models reach ≈99–100% accuracy; in mid tiers (A10–A16) linear SVM leads with RF100 and RF125 close behind; and in the weakest tiers (A22–A25) all models deteriorate, with k-NN keeping a small edge over RF while SVM drops more sharply. Figure 4 shows that these patterns are stable across the three test sizes: Figure 4(a) 10%, Figure 4(b) 25%, and Figure 4(c) 40% test sizes; differences among the test sizes are minor relative to the effect of correlation. Figure 4 summarizes the trajectories across tiers and test sizes. Full 10% and 40% tables and metric panels are provided in the repository [59].

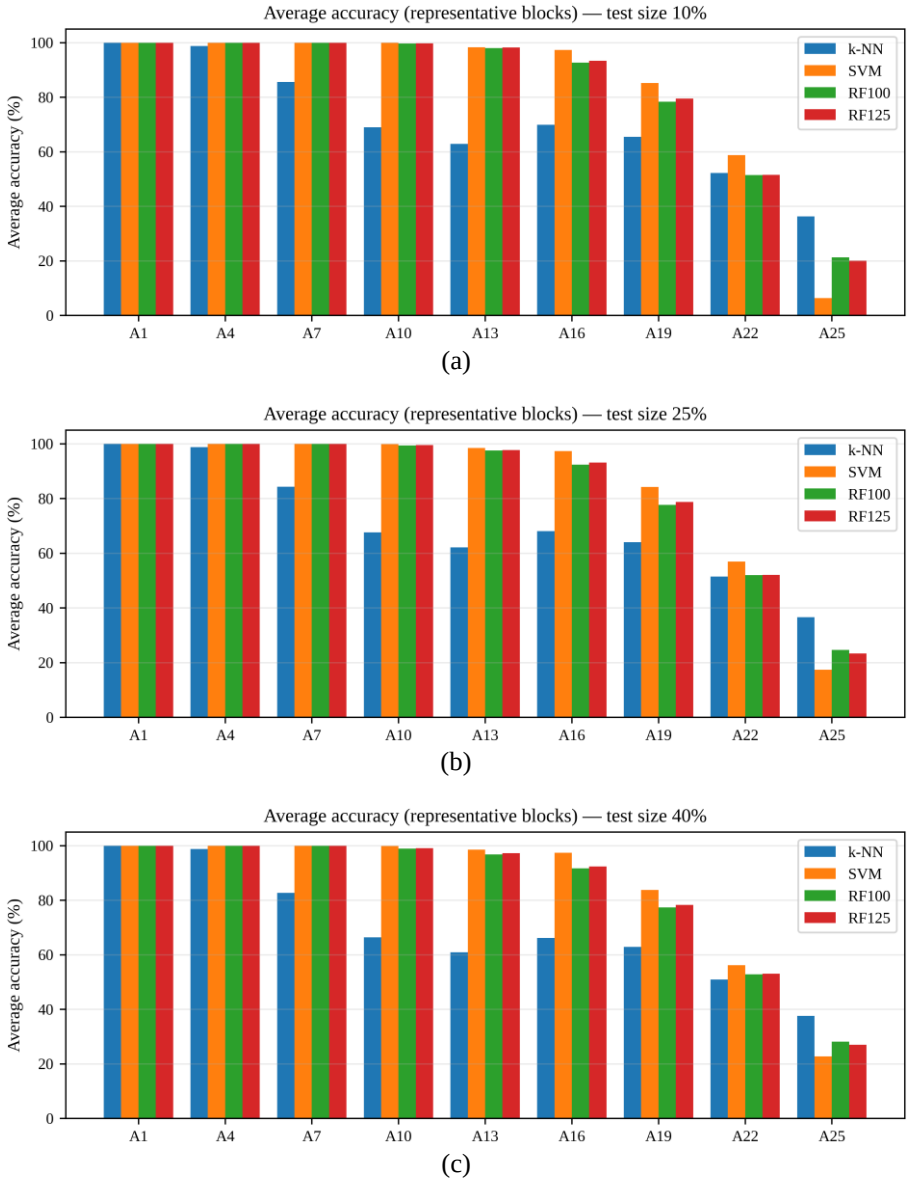


Figure 4. Average accuracy across representative correlation blocks for three test sizes; (a) 10%, (b) 25%, and (c) 40% (classifiers: k-NN, SVM, RF100, and RF125)

4.1.2. Receiver operating characteristic–area under the curve and error profiles

The mid-correlation block A16 serves as the representative case in the body. Figure 5 reports ROC curves on A16 for the 10%, 25%, and 40% test sizes (diagonal denotes chance), and Figure 6 reports the row-normalized confusion matrix on A16 at the 25% test size. On A16, SVM attains the highest AUC (≈1), RF100 and RF125 are marginally lower, and k-NN remains competitive but non-dominant; differences among test sizes are minor relative to correlation strength. At higher-correlation tiers (e.g., A7), ROC traces lie near the upper-left corner and CM show dominant diagonals; at the weakest tier (e.g., A25), ROC curves drift toward the diagonal and the matrix collapses toward a single class. Full ROC and CM grids for other blocks and test sizes are provided in the repository [59].

Overall, DatasetA shows a clear correlation-driven difficulty gradient with SVM most resilient, RF a close second, and k-NN competitive only at the extremes, a pattern stable across MCCV resamples and test sizes. Extended heatmaps and boxplots for DatasetA tiers are available in the repository [59].

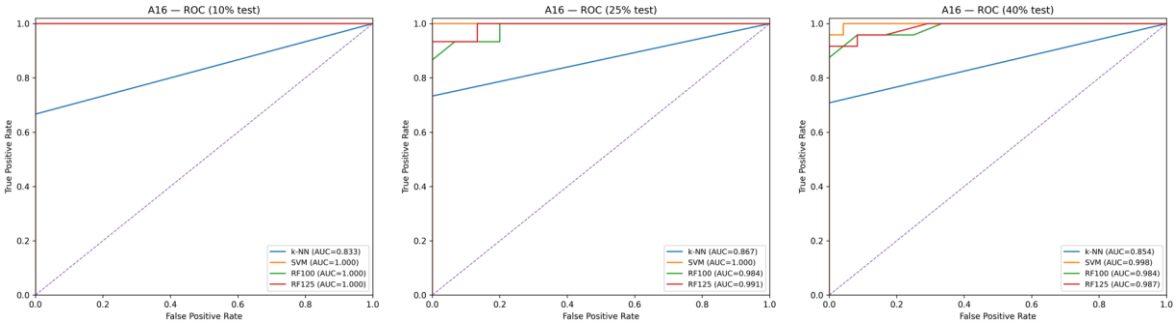


Figure 5. ROC curves on A16 (moderate correlation) for 10%, 25%, and 40% test sizes; k-NN, linear SVM, RF100, RF125 (legend/linestyle are consistent; color-blind-safe variants are provided in [59])

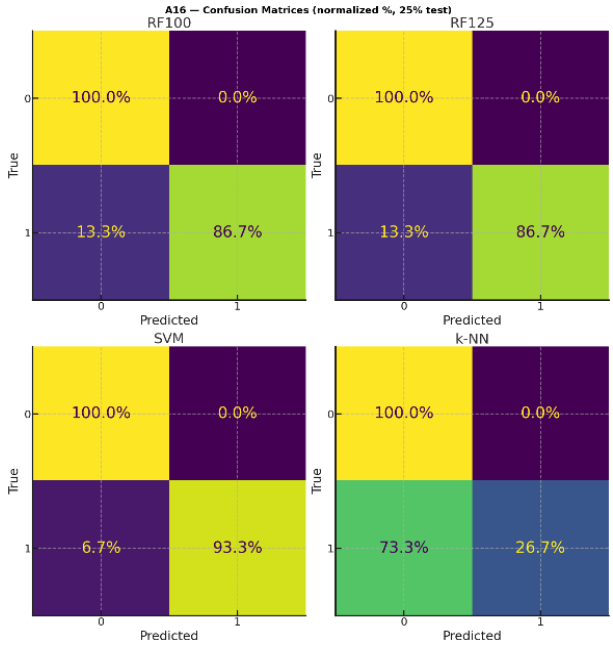


Figure 6. CM on A16 (25% test size; normalized, values in %)

4.2. Performance on DatasetB (negatively correlated single-nucleotide polymorphisms)

This section reports DatasetB (negative-correlation) results by correlation tier and test size. Representative results at 25% test size are presented in Table 6 for eight tiers (B1, B4, B7, B10, B13, B16, B19, B22). Complete summaries for the 10% and 40% test sizes, full ROC and confusion-matrix (CM) panels, AUC tables, and extended heatmaps/boxplots for DatasetB tiers are available in the repository [59].

Table 6. DatasetB: average accuracy (%) of RF, k-NN, and SVM with 25% test data

Pearson corr. range	B1	B4	B7	B10	B13	B16	B19	B22
k-NN	34.39	48.47	61.18	66.13	71.48	91.59	98.5	100
SVM	17.09	47.58	84.18	98.5	100	100	100	100
RF100	24.21	48.21	75.84	93.41	99.4	99.95	100	100
RF125	22.87	48.45	75.89	93.54	99.42	99.99	100	100

4.2.1. Accuracy and metric trends

Accuracy increases with the magnitude of negative correlation: performance is low at B1, improves rapidly through B4–B10, and approaches ceiling by B19–B22 (≈99–100% across models; at B16, SVM/RF are ≈100% while k-NN ≈92%). Differences among classifiers are most visible in weakly correlated tiers; in mid–strong tiers all methods converge around ≈98–100% accuracy. The same pattern holds at 10%, /25%, and 40% test sizes; full summaries are available in the repository [59].

4.2.2. Receiver operating characteristic-area under the curve and error profiles

Figure 7 reports ROC curves on B10 (−0.285 to −0.254) at the 25% test size: all models achieve high discriminative ability (AUCs>0.97), with linear SVM=0.999, RF125=0.989, RF100=0.987, and k-NN=0.973. Figure 8 presents row-normalized CM (25%) for B1 (low), B10 (mid), and B22 (high): frequent misclassifications at B1, sharply reduced errors at B10, and perfect separation at B22. These visuals, together with Table 6, indicate that correlation magnitude—not its sign—governs separability. Full panels and AUC tables are provided in the repository cited above.

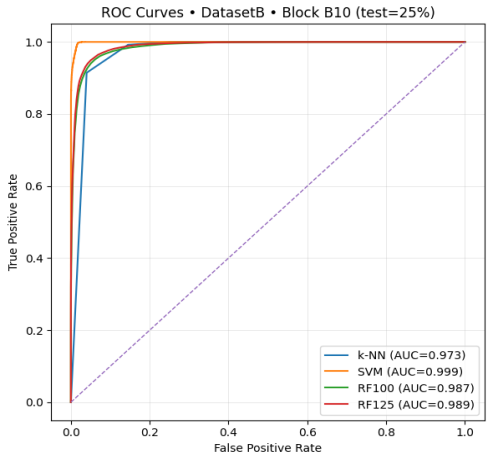


Figure 7. ROC curves of k-NN, SVM, RF100, and RF125 on DatasetB (B10, −0.285 to −0.254; 25% test) (legend/linestyle are consistent; color-blind-safe variants are provided in [59])

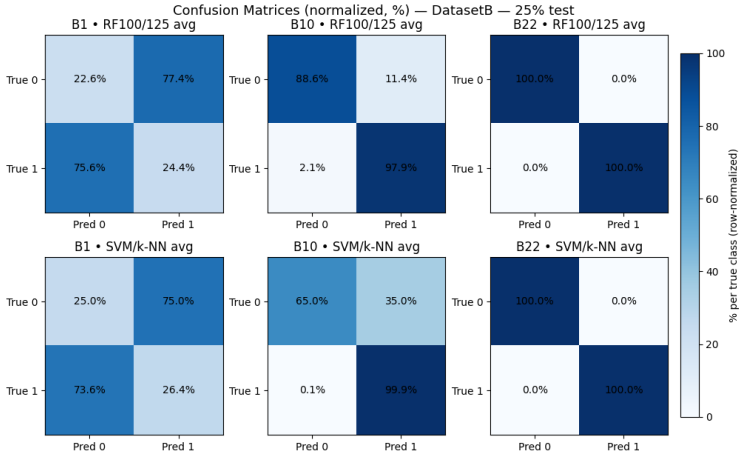


Figure 8. CM (normalized, %) for DatasetB (25% test) at low (B1), mid (B10), and high (B22) correlation blocks

4.3. Significance testing across classifiers

Shapiro–Wilk tests reject normality in most moderate–weak blocks ($p \leq 0.05$) and show mixed outcomes in strong blocks; accordingly, nonparametric inference is used. Kruskal–Wallis omnibus tests are significant across all examined blocks and test sizes ($p < 0.05$), indicating at least one classifier differs in every condition. Selected summaries of Shapiro–Wilk normality checks and Kruskal–Wallis omnibus tests are provided for representative weak, moderate, and strong blocks—A25/A16/A7 (DatasetA) and B1/B10/B19 (DatasetB)—at 10%, 25%, and 40% test sizes. These materials are available in the repository [59].

4.3.1. Normality checks (Shapiro–Wilk)

Shapiro–Wilk tests ($\alpha = 0.05$) on the distributions of 1,000 MCCV accuracies largely reject normality in moderate–weak blocks (A16, A25, B10, and B1) across 10%, 25%, and 40%, while strong blocks (A7, B19) show mixed outcomes: SVM (and often RF125) tends not to reject normality, whereas k-NN (and sometimes RF100) often remains non-normal. Accordingly, subsequent inference uses nonparametric procedures.

4.3.2. Omnibus differences (Kruskal–Wallis)

Kruskal–Wallis omnibus tests ($\alpha = 0.05$) across six representative blocks (A7, A16, A25, B1, B10, and B19) and three test sizes (10%, 25%, and 40%) are significant in all conditions ($p < 0.05$), indicating that at least one classifier differs in every block–size combination.

4.3.3. Pairwise contrasts (Mann–Whitney with Bonferroni)

Bonferroni-adjusted Mann–Whitney tests confirm pairwise differences that mirror the accuracy/ROC patterns. Across test sizes (Table 7), RF100 vs RF125 shows few differences (significant in 1/6 contrasts at 10%, 3/6 at 25% and 40%; median $|\delta| \approx 0.09$). SVM vs RF pairs are significant in 4/6 contrasts at all sizes (median $|\delta| \approx 0.53$–0.66). k-NN vs RF is significant in 6/6 contrasts with large effects (median $|\delta| \approx 0.75$–0.90), and k-NN vs SVM is also significant in 6/6 contrasts with very large effects (median $|\delta| \approx 0.95$–0.97); directions follow the tier-wise patterns: at strong correlation (e.g., A7, B19) ceiling effects render SVM vs RF often not significant, at mid correlation (e.g., A16) SVM exceeds RF, and at weak correlation (e.g., A25) k-NN can exceed RF and may exceed SVM. A compact cross–test-size summary appears in Table 7; complete pairwise tables for A7, A16, A25 (DatasetA) and B1, B10, B19 (DatasetB)—including confidence intervals and adjusted p-values—are available in [59].

Table 7. Summary of pairwise Mann–Whitney U results across test sizes (10%, 25%, and 40%)

Model pair	10% test size		25% test size		40% test size	
	#significant/6	Median |δ|	#significant/6	Median |δ|	#significant/6	Median |δ|
RF100 vs RF125	1	0.091	3	0.094	3	0.108
SVM vs RF100	4	0.554	4	0.600	4	0.660
SVM vs RF125	4	0.532	4	0.526	4	0.588
k-NN vs RF100	6	0.751	6	0.868	6	0.870
k-NN vs RF125	6	0.781	6	0.889	6	0.900
k-NN vs SVM	6	0.950	6	0.972	6	0.963

4.3.4. Summary and implications

Classifier choice materially affects accuracy: SVM is consistently strongest from mid to high correlation ranges; RF trails closely with minimal sensitivity to tree count (RF100 vs RF125 significant in 1/6, 3/6, 3/6 contrasts; median $|\delta| \approx 0.09$; Table 7); k-NN is competitive at the strongest tiers and occasionally superior in the weakest tiers, but generally dominated elsewhere. These patterns hold across DatasetA/B and the 10%, 25%, 40% test sizes (omnibus Kruskal–Wallis $p < 0.05$).

4.4. Pearson (correlation-ranked) vs PCA-120 baseline

This subsection benchmarks a correlation-ranked panel (“Pearson-120”) against a label-free projection baseline (“PCA-120”) and assesses whether any differences persist across classifiers and the 10%, 25%, and 40% test sizes.

4.4.1. Feature-panel construction and evaluation protocol

This subsection contrasts Pearson-based selection with a label-free projection baseline (PCA-120). PCA was fitted within each MCCV repetition on TRAIN folds only, after standardizing genotypes, to prevent information leakage. From the post-QC HapMap Phase II matrix (120×9305; dosage 0/1/2; TRAIN-fold minor-allele mapping and median imputation), loci were ranked by EVR-weighted loading magnitudes, and the top 120

were retained (PCA-120, DatasetA+DatasetB). For a matched-dimensionality comparator, Pearson-120 denotes the 120 SNPs with the largest absolute Pearson correlation (|r|) with the label from the strongest block (A1).

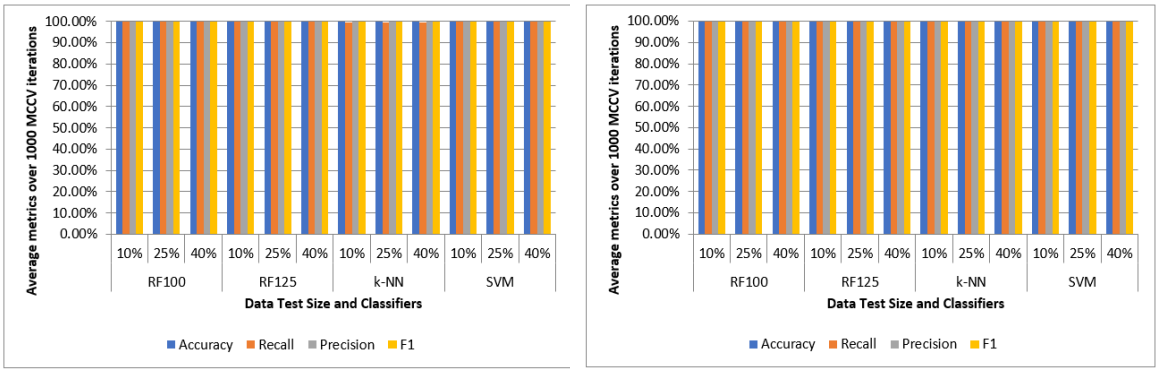
4.4.2. Results

Across R=1,000 MCCV resamples at 10%, 25%, and 40% test sizes, SVM and RF100/RF125 achieve ≈99–100% for accuracy, recall, precision, and F1 under both Pearson-120 and-120 (DatasetA+DatasetB). The only non-ceiling pattern appears for k-NN at 10% with PCA-120, where mean recall and F1 are ≈0.2 percentage points below 100%; at 25% and 40%, k-NN returns to ≈100%. Under Pearson-120, all four classifiers reach 100% for accuracy, recall, precision, and F1 at all test sizes. Table 8 summarizes PCA-120 (DatasetA+DatasetB) and Figure 9 juxtaposes PCA-120 (panel a) vs Pearson-120 (panel b).

Table 8. PCA-120 (DatasetA+DatasetB): mean accuracy, recall, precision, and F1 (%) across classifiers at the 10%, 25%, and 40% test sizes (MCCV=1000)

Classifier	RF100			RF125			k-NN			SVM		
Test size	10%	25%	40%	10%	25%	40%	10%	25%	40%	10%	25%	40%
Acc. (%)	99.99	99.97	99.95	100	99.98	99.98	99.79	99.79	99.75	100	100	100
Recall (%)	99.98	99.98	99.92	100	99.98	99.96	99.58	99.58	99.51	100	100	100
Prec. (%)	100	99.97	99.99	100	99.98	100	100	100	100	100	100	100
F1 (%)	99.99	99.97	99.95	100	99.98	99.98	99.77	99.78	99.75	100	100	100

Figure 9(a) shows the PCA-120 results (DatasetA+DatasetB, no labels used in selection), whereas Figure 9(b) shows the Pearson-120 results; both panels report mean performance over 1,000 MCCV resamples at the 10%, 25%, and 40% test sizes



(a) (b)

Figure 9. Performance with 120 features; (a) PCA-120 (DatasetA+DatasetB) and (b) Pearson-120; means over 1,000 MCCV at 10%, 25%, and 40% test sizes

4.4.3. Summary and implications

Both strategies yield ≈99–100% accuracy for SVM and RF, while Pearson-based selection retains a small advantage for k-NN at the 10% test size. In this sense, the methods are complementary: Pearson-based selection preserves neighborhood structure that benefits distance-based classifiers, whereas PCA provides a label-independent baseline with comparable ceiling performance for margin-based and ensemble models. Complete numerical tables for the Pearson-120 and PCA-120 comparisons—including per-iteration metrics, per-test-size aggregates, selected SNP panels, and MCCV selection frequencies—are available in the repository [59].

4.5. Runtime and computational footprint

Runtime profiles corroborate deployability: with 120-SNP panels, prediction latencies are millisecond-scale for linear SVM and k-NN and sub-tenth-second for RF, with fit+predict totals summarized in Table 9. Per-iteration timings (median [IQR], ms) across test sizes show that, at the 25% test size, median fit costs follow k-NN (0.74–1.00 ms) < SVM (≈2.00 ms) ≪ RF100 (231–233 ms) < RF125 (269–272 ms); median prediction costs are SVM (≈0.30 ms), k-NN (1.81–1.83 ms), and RF (34–35 ms). Consequently, median

total time per iteration is SVM (2.21–2.36 ms) < k-NN (2.49–2.53 ms) ≪ RF100 (267–273 ms) < RF125 (312–313 ms), with narrow IQRs indicating stable runtimes over R=1,000 MCCV resamples. Overall, runtimes are modest, demonstrating that both the Pearson-based selection pipeline and the PCA baseline are computationally practical.

Table 9. Per-iteration runtime (milliseconds), reported as median and IQR, for each classifier at 10%, 25%, and 40% test sizes (timing includes model fit() and predict() only)

Classifier	Test size (%)	Median fit	IQR fit	Median prediction	IQR prediction	Median total	IQR total
RF100	10	232.063	29.302	34.57	0.992	266.906	29.531
RF100	25	231.481	36.419	34.582	1.088	266.397	33.963
RF100	40	238.567	23.624	34.587	0.949	273.465	19.661
RF125	10	272.905	23.886	34.873	2.751	311.518	28.365
RF125	25	272.363	23.496	35.045	10.345	312.051	26.315
RF125	40	271.495	23.518	35.221	10.593	312.503	26.984
SVM	10	2.071	0.504	0.3	0.079	2.364	0.579
SVM	25	1.978	0.505	0.298	0.083	2.266	0.595
SVM	40	1.92	0.512	0.304	0.087	2.213	0.588
k-NN	10	0.752	0.163	1.766	0.358	2.493	0.512
k-NN	25	0.737	0.217	1.81	0.374	2.525	0.605
k-NN	40	0.717	0.218	1.829	0.4	2.522	0.617

4.5.1. Practical implication

Given the near-parity in accuracy at 120 features (§4.4), SVM offers the best accuracy–latency trade-off; k-NN is close behind but incurs higher prediction cost; RF variants are ≈100× (about two orders of magnitude) slower for both training and inference, so runtime can guide model choice in resource-constrained deployments. Complete per-iteration runtime traces supporting Table 9—covering fit, predict, and total times for each model across the 10%, 25%, and 40% test sizes and all representative blocks—are available in the repository [59].

4.6. Consensus panel

A consensus SNP panel was derived by combining two signals—the percentile of absolute point-biserial correlation (|r|) with the label and the percentile of PCA selection frequency across MCCV resamples. The combined score yields nested panels (Top-50, Top-120, Top-240); Top-120 balances parsimony and stability and is used as the primary panel. The Top-50, Top-120, Top-240, and the full consensus panel are available in the public repository [59].

4.6.1. Top-20 minimal panel

For operational use and cost-sensitive assays, a Top-20 subset of the consensus panel is reported as an illustrative, low-complexity option (Table 10). This subset preserves coverage of the highest $|r|$ percentiles and, in mid-to-high correlation tiers, attains ≈98–100% accuracy for SVM/RF, with only small deltas relative to Top-120; per-tier accuracy differences and confidence intervals are provided in [59].

Table 10. Top-20 consensus SNPs ranked by combined correlation and PCA-selection percentiles

Feature	|r|	Pearson_percentile	Select_rate overall	PCA percentile	Combined score
rs10868791	0.925	100	0	92.667	96.334
rs6670842	0.900	99.982	0	92.667	96.325
rs9909962	0.900	99.965	0	92.667	96.316
rs2370893	0.900	99.947	0	92.667	96.307
rs6814827	0.897	99.929	0	92.667	96.298
rs311992	0.874	99.911	0	92.667	96.289
rs10504132	0.873	99.894	0	92.667	96.281
rs13420968	0.869	99.876	0	92.667	96.272
rs9534610	0.868	99.858	0	92.667	96.263
rs1485768	0.867	99.841	0	92.667	96.254
rs7752055	0.863	99.823	0	92.667	96.245
rs1209914	0.849	99.805	0	92.667	96.236
rs1373013	0.847	99.788	0	92.667	96.227
rs2034510	0.843	99.770	0	92.667	96.219
rs1568773	0.840	99.752	0	92.667	96.210
rs619228	0.835	99.734	0	92.667	96.201
rs2833795	0.834	99.717	0	92.667	96.192
rs7851392	0.834	99.699	0	92.667	96.183
rs6716734	0.828	99.681	0	92.667	96.174
rs2003154	0.828	99.664	0	92.667	96.165

4.6.2. Correlation thresholds for target accuracies

Block-wise accuracy across DatasetA and DatasetB supports interpretable, label-aware thresholds for marker prioritization (R=1,000 MCCV resamples; 10%, 25%, and 40% test sizes). Using k-NN as the conservative yardstick, |r| ≥0.50 attains ≥90%, |r| ≥0.52 achieve ≥95%, and |r| ≥0.60 reaches ≈99%; classifier-specific minima are summarized in Table 11. These thresholds complement the consensus panel when rank ties or assay constraints must be resolved.

Table 11. Minimum absolute correlation (|r|min) required to reach ≥90%, ≥95%, and ≈99% accuracy across datasets (A and B) and classifiers (global thresholds are conservatively determined by k-NN)

Classifier	≥ 90% (|r|min)	≥ 95% (|r|min)	≈ 99% (|r|min)
k-NN	0.5	0.52	0.6
SVM	0.44	0.44	0.44
RF100	0.44	0.44	0.44
RF125	0.44	0.44	0.44
Global (all models)	0.5	0.52	0.6

4.6.3. Summary and practical guidance

For discovery and flexible downstream analysis, Top-120 offers a stable, interpretable panel with ≈99–100% performance in mid–high correlation tiers; for rapid deployments or budget-limited wet-lab follow-up, Top-20 is an attractive default with ≈98–100% in those tiers and small deltas relative to Top-120 (per-tier summaries in [59]). When assay slots are scarce, the $|r|_{\min}$ thresholds in Table 11 (0.50/0.52/0.60 for ≥90%/≥95%/≈99%) provide a simple rule to finalize inclusions or substitutions while maintaining target accuracy.

4.7. Overall discussion and practical implications

Results across DatasetA and DatasetB and all test sizes show consistent trends: classifier performance scales with the strength of feature–label correlation, SVM and RF reach near-ceiling accuracy (≈98–100%; e.g., A1–A7 and B16–B22) in moderate-to-strong blocks, and k-NN degrades more steeply as correlation weakens. Statistical validation indicates that these differences are statistically significant and consistent across blocks, reinforcing that feature–label correlation is the key determinant of predictive accuracy.

In mid-correlation tiers, linear SVM's advantage is consistent with margin-based generalization: once |r| supplies moderately informative axes, a linear separator attains large, stable margins and low variance. RF approaches ceiling as correlation strengthens but is slightly more variance-prone in the mid-range due to tree-split instability on weaker signals. k-NN relies on local neighborhood purity; it benefits most at very high |r| yet degrades faster as manifolds overlap or noise increases. These mechanisms mirror the observed accuracy and pairwise-significance patterns across DatasetA/B blocks and test sizes.

Methodologically, a simple Pearson-based selection yields performance competitive with more complex pipelines. Relative to the PCA baseline, Pearson-based selection achieves indistinguishable accuracy for SVM and RF, and slightly outperforms PCA for k-NN at smaller test sizes—underscoring the practical value of a transparent, interpretable selector while still recognizing PCA as a strong unsupervised reference.

Runtime analysis (Table 9) shows that RF carries the largest training cost (RF125 > RF100), SVM is consistently fast, and k-NN is negligible at training but heavier at prediction. Median per-iteration runtimes remain modest—hundreds of milliseconds for RF and only a few milliseconds for SVM and k-NN—demonstrating that both the Pearson-based selection pipeline and the PCA baseline are computationally practical.

4.7.1. CPU-only feasibility

All experiments ran on CPU. Under 120-SNP panels, linear SVM/k-NN operate in milliseconds per query, while RF remains sub-second, and fit overheads follow SVM ≪ RF100 < RF125 (Table 9). These latencies, consistent across R=1000 MCCV resamples and all test sizes, indicate practicality for point-of-care or edge deployment.

4.7.2. Several limitations merit note

LD-aware redundancy control (e.g., LD-clumping) was not systematically applied; only two HapMap Phase II populations (CEU vs YRI) were analyzed; and the sample size is relatively small. Future work will incorporate LD-clumping or mRMR to reduce redundancy, expand to multi-ethnic and multi-class cohorts, and evaluate deep representation learning and functional-annotation integration, while preserving leakage-free training protocols and locus-level interpretability.

Taken together—i) parity with PCA-120, ii) stable runtimes at fixed dimensionality, iii) a compact Top-20 with near-ceiling performance, and iv) actionable |r| thresholds—these findings support Pearson-based selection as a simple, transparent baseline for small-n/high-p SNP classification, with the consensus construction adding robustness while remaining block-agnostic.

5. CONCLUSION

Pearson-based selection, complemented by a simple consensus rule, delivers interpretable panels that match PCA-120 performance at the same dimensionality while simplifying deployment decisions. A 20-locus practical panel provides a cost-effective option with minimal accuracy loss, and conservative |r| thresholds translate results into clear operating points. These outcomes recommend Pearson-ranked, consensus-refined subsets as a strong baseline for SNP-based classification under small-n/high-p constraints.

Under systematic QC and Pearson-based selection, supervised classification of human populations shows robust, statistically significant differences among classifiers: RF and SVM reach ceiling performance on moderate-to-strong correlation subsets, whereas k-NN is more sensitive as correlation weakens. Compared with PCA, Pearson-based selection remains competitive—SVM and RF attain indistinguishable accuracy under both, and k-NN retains a small edge with Pearson-guided features—while runtime demands are modest.

Beyond accuracy, two practical outputs are provided: i) a 120-SNP consensus panel that combines correlation evidence with PCA selection frequency, and ii) operating thresholds ($|r| \geq 0.50$ for ≥90%, $|r| \geq 0.52$ for ≥95%, $|r| \geq 0.60$ for ≈99%) that offer reproducible criteria for marker prioritization.

Limitations include the absence of LD-clumping, restriction to two HapMap Phase II populations, and limited sample size; future work should broaden cohorts, incorporate LD-aware pruning and functional annotations, and assess stability at larger scales.

On commodity CPUs, inference with 120-SNP panels is millisecond-scale for linear SVM/k-NN and sub-second for RF, enabling point-of-care or edge deployment. Future work will extend to multi-class cohorts and deep representation baselines under the same leakage-free protocol, while maintaining locus-level interpretability. These steps operationalize a reproducible, CPU-feasible pipeline for real-world screening scenarios.

ACKNOWLEDGMENTS

This research was supported by Universitas Kristen Satya Wacana through the internal research scheme in 2023. We are grateful for their contribution to this research. Additionally, we would like to express our appreciation to the anonymous reviewers for their valuable feedback that significantly enhanced the quality of our paper.

FUNDING INFORMATION

This research was supported by Universitas Kristen Satya Wacana through the internal research scheme in 2023 with Decree No. 201/RIK-RPM/9/2023.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Prihanto Ngesti Basuki	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Sri Yulianto Joko Prasetyo	✓	✓			✓	✓		✓		✓	✓	✓		
Adi Setiawan	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no financial, personal, or professional conflicts of interest that could have influenced the work reported in this paper.

INFORMED CONSENT

Not applicable. This study used publicly available, de-identified genotype data and did not involve human participants recruitment.

ETHICAL APPROVAL

Not applicable. This study analyzed publicly available, de-identified data and did not require institutional ethical approval.

DATA AVAILABILITY

All datasets, analysis scripts, and extended tables/figures are openly available in the repository [59] (concept DOI), ensuring exact reproducibility of the reported results.

REFERENCES

[1] Z. Ahmed, S. Zeeshan, D. Mendhe, and X. Dong, "Human gene and disease associations for clinical-genomics and precision medicine research," *Clinical and Translational Medicine*, vol. 10, no. 1, pp. 297–318, 2020, doi: 10.1002/ctm2.28.

[2] C. Fabbri, "Genetics in psychiatry: Methods, clinical applications and future perspectives," *Psychiatry and Clinical Neurosciences Reports*, vol. 1, no. 2, pp. 1-13, Jun. 2022, doi: 10.1002/pcn5.6.

[3] P. Guevara-Ramírez *et al.*, "Genetics, genomics, and diet interactions in obesity in the Latin American environment," *Frontiers MediaFfig*, vol. 9, pp. 1-21, 2022, doi: 10.3389/fnut.2022.1063286.

[4] S. Lip and S. Padmanabhan, "Genomics of Blood Pressure and Hypertension: Extending the Mosaic Theory Toward Stratification," *Canadian Journal of Cardiology*, vol. 36, no. 5, pp. 694–705, May 2020, doi: 10.1016/j.cjca.2020.03.001.

[5] Z. Jia *et al.*, "Single nucleotide polymorphisms associated with female breast cancer susceptibility in Chinese population," *Gene*, vol. 884, p. 147676, Feb. 2023, doi: 10.1016/j.gene.2023.147676.

[6] S. Kaur, A. Ali, U. Ahmad, Y. Siahbalaei, A. K. Pandey, and B. Singh, "Role of single nucleotide polymorphisms (SNPs) in common migraine," *Egyptian Journal of Neurology, Psychiatry and Neurosurgery*, vol. 55, no. 1, pp. 1-7, Dec. 2019, doi: 10.1186/s41983-019-0093-8.

[7] I. Hayah, C. Talbi, N. Chafai, I. Houaga, S. Botti, and B. Badaoui, "Genetic diversity and breed-informative SNPs identification in domestic pig populations using coding SNPs," *Frontiers in Genetics*, vol. 14, pp. 1-11, 2023, doi: 10.3389/fgene.2023.1229741.

[8] J. Ruiz-Ramírez *et al.*, "Development and evaluations of the ancestry informative markers of the VISAGE Enhanced Tool for Appearance and Ancestry," *Forensic Science International Genetics*, vol. 64, no. 1, pp. 1-23, May 2023, doi: 10.1016/j.fsigen.2023.102853.

[9] X. Y. Jin *et al.*, "Biogeographic origin prediction of three continental populations through 42 ancestry informative SNPs," *Electrophoresis*, vol. 41, no. 3–4, pp. 235–245, Feb. 2020, doi: 10.1002/elps.201900241.

[10] S. Sehrawat, K. Najafian, and L. Jin, "Predicting phenotypes from novel genomic markers using deep learning," *Bioinformatics Advances*, vol. 3, no. 1, 2023, doi: 10.1093/bioadv/vbad028.

[11] M. N. Murty and M. Avinash, *Representation in Machine Learning*, Springer Singapore, 2023, doi: 10.1007/978-981-19-7908-8.

[12] N. Kalcheva, M. Todorova, and I. Penev, "Study of the K-Nearest Neighbors Method with Various Features for Text Classification in Machine Learning," in *2023 International Conference Automatics and Informatics (ICAI)*, 2023, pp. 37–40, doi: 10.1109/ICAI58806.2023.10339061.

[13] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for Machine Learning*. Cambridge: Cambridge University Press, 2020, doi: 10.1017/9781108679930.

[14] J. M. Nguyen *et al.*, "Random forest of perfect trees: concept, performance, applications and perspectives," *Bioinformatics*, vol. 37, no. 15, pp. 2165–2174, 2021, doi: 10.1093/bioinformatics/btab074.

[15] R. Devika, S. V. Avilala, and V. Subramaniyaswamy, "Comparative study of classifier for chronic kidney disease prediction using naive bayes, KNN and random forest," in *Proceedings of the 3rd International Conference on Computing Methodologies and Communication (ICCMC) 2019*, 2019, pp. 679–684, doi: 10.1109/ICCMC.2019.8819654.

[16] I. Mohit, K. S. Kumar, A. U. K. Reddy, and B. S. Kumar, "An Approach to detect multiple diseases using machine learning algorithm," *1st International Conference on Applied Mathematics, Modeling and Simulation in Engineering (AMSE) 2021*, 2021, vol. 2089, no. 1, doi: 10.1088/1742-6596/2089/1/012009.

[17] C. Gupta, A. Saha, N. V. S. Reddy, and U. D. Acharya, "Cardiac Disease Prediction using Supervised Machine Learning Techniques," in *1st International Conference on Artificial Intelligence, Computational Electronics and Communication System (AICECS 2021)*, 2022, vol. 2161, no. 1, doi: 10.1088/1742-6596/2161/1/012013.

[18] K. Veena, K. Meena, Y. Teekaraman, R. Kuppusamy, and A. Radhakrishnan, "C SVM Classification and KNN Techniques for Cyber Crime Detection," *Wireless Communications and Mobile Computing*, vol. 2022, 2022, doi: 10.1155/2022/3640017.

[19] M. R. Astari, M. T. Nuruzzaman, and B. Sugiantoro, "Comparison of K-Nearest Neighbor, Support Vector Machine, Random Forest, and C 4.5 Algorithms on Indoor Positioning System," *IJID (International Journal on Informatics for Development)*, vol. 12, no. 1, pp. 302–313, 2023, doi: 10.14421/ijid.2023.3991.

[20] M. O. Gani, R. K. Ayyasamy, A. Sangodiah, and Y. T. Fui, "USTW Vs. STW: A Comparative Analysis for Exam Question Classification based on Bloom's Taxonomy," *MENDEL-Soft Computing Journal*, vol. 2, pp. 2571–3701, 2022, doi: 10.13164/mendel.202..025.

[21] X. Ma *et al.*, "Comparison and development of machine learning tools for the prediction of chronic obstructive pulmonary disease in the Chinese population," *Journal of Translational Medicine*, vol. 18, no. 1, pp. 1-14, Mar. 2020, doi: 10.1186/s12967-020-02312-0.
[22] D. O. Enoma, J. Bishung, T. Abiodun, O. Ogunlana, and V. C. Osamor, "Machine learning approaches to genome-wide association studies," *Journal of King Saud University - Science*, vol. 34, no. 4, pp. 1-9, 2022, doi: 10.1016/j.jksus.2022.101847.
[23] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," *Frontiers in Bioinformatics*, vol. 2, pp. 1–17, 2022, doi: 10.3389/fbinf.2022.927312.
[24] K. E. Grinde, B. L. Browning, A. P. Reiner, T. A. Thornton, and S. R. Browning, "Adjusting for principal components can induce collider bias in genome-wide association studies," *PLoS Genetics*, vol. 20, no. 12, pp. 1–29, 2024, doi: 10.1371/journal.pgen.1011242.
[25] J. Van Waaij, S. Li, G. Garcia-Erill, A. Albrechtsen, and C. Wiuf, "Evaluation of population structure inferred by principal component analysis or the admixture model," *Genetics*, vol. 225, no. 2, pp. 1-15, 2023, doi: 10.1093/genetics/iyad157.
[26] M. K. M. Rabby, A. K. M. K. Islam, S. Belkasim, and M. U. Bikdash, "Epileptic seizures classification in EEG using PCA based genetic algorithm through machine learning," in *Proceedings of the 2021 ACMSE Conference - ACMSE 2021: The Annual ACM Southeast Conference*, 2021, pp. 17–24, doi: 10.1145/3409334.3452065.
[27] Z. Liu, I. Barnett, and X. Lin, "A comparison of principal component methods between multiple phenotype regression and multiple SNP regression in genetic association studies," *Annals of Applied Statistics*, vol. 14, no. 1, pp. 433–451, 2020, doi: 10.1214/19-AOAS1312.
[28] E. Elhaik, "Why most Principal Component Analyses (PCA) in population genetic studies are wrong," *bioRxiv*, 2021, doi: 10.1101/2021.04.11.439381.
[29] Y. Yao and A. Ochoa, "Limitations of principal components in quantitative genetic association models for human studies," *Elife*, vol. 12, pp. 1–36, 2023, doi: 10.7554/eLife.79238.
[30] E. Elhaik, "Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated," *Scientific Reports*, vol. 12, no. 1, pp. 1–35, 2022, doi: 10.1038/s41598-022-14395-4.
[31] J. Ahlinder, D. Hall, M. Suontama, and M. J. Sillanpää, "Principal component analysis revisited: fast multitrait genetic evaluations with smooth convergence," *G3 (Bethesda)*, vol. 14, no. 12, pp. 1-17, 2024, doi: 10.1093/g3journal/jkae228.
[32] H. Gong, Y. Li, J. Zhang, B. Zhang, and X. A. Wang, "A new filter feature selection algorithm by ensembling Pearson correlation coefficient and mutual information," *Engineering Applications of Artificial Intelligence*, vol. 131, 2024, doi: 10.1016/j.engappai.2024.107865.
[33] B. Kalaiselvi and M. Thangamani, "An efficient Pearson correlation based improved random forest classification for protein structure prediction techniques," *Measurement (Lond)*, vol. 162, p. 107885, 2020, doi: 10.1016/j.measurement.2020.107885.
[34] S. Kumar, B. Bhushan, L. Bhambhu, M. Thakur, U. M. Mohapatra, and D. K. Choubey, "Medical Datasets Classification using a Hybrid Genetic Algorithm for Feature Selection based on Pearson Correlation Coefficient," in *2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS)*, 2022, pp. 214–218, doi: 10.1109/MLCSS57186.2022.00047.
[35] P. N. Basuki, J. P. S. Yulianto, and A. Setiawan, "Comparison of KNN and SVM Methods for the Accuracy of Individual Race Classification Prediction Based on SNP Genetic Data," in *Proceeding of the 3rd International Conference on Electronics, Biomedical Engineering, and Health Informatics*, Eds., Singapore: Springer Nature Singapore, 2023, pp. 411–427, doi: 10.1007/978-981-99-0248-4_28.
[36] G. Shan, "Monte Carlo cross-validation for a study with binary outcome and limited sample size," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1–15, 2022, doi: 10.1186/s12911-022-02016-z.
[37] F. Macedo, R. Valadas, E. Carrasquinha, M. R. Oliveira, and A. Pacheco, "Feature selection using Decomposed Mutual Information Maximization," *Neurocomputing*, vol. 513, pp. 215–232, 2022, doi: 10.1016/j.neucom.2022.09.101.
[38] J. Mielniczuk, "Information Theoretic Methods for Variable Selection—A Review," *Entropy*, vol. 24, no. 8, pp. 1–25, 2022, doi: 10.3390/e24081079.
[39] H. Hamla and K. Ghanem, "A Hybrid Feature Selection Based on Fisher Score and SVM-RFE for Microarray Data," *Informatica (Slovenia)*, vol. 48, no. 1, pp. 57–68, 2024, doi: 10.31449/inf.v48i1.4759.
[40] O. Bulut, B. Tan, E. Mazzullo, and A. Syed, "Benchmarking Variants of Recursive Feature Elimination: Insights from Predictive Tasks in Education and Healthcare," *Information*, vol. 16, no. 6, pp. 1–21, 2025, doi: 10.3390/info16060476.
[41] J. Crawford, M. Chikina, and C. S. Greene, "Optimizer's dilemma: optimization strongly influences model selection in transcriptomic prediction," *Bioinformatics Advances*, vol. 4, no. 1, pp. 1-8, 2024, doi: 10.1093/bioadv/vbae004.
[42] S. G. Feronato *et al.*, "Selecting Genetic Variants and Interactions Associated with Amyotrophic Lateral Sclerosis: A Group LASSO Approach," *Journal of Personalized Medicine*, vol. 12, no. 8, pp. 1-18, 2022, doi: 10.3390/jpm12081330.
[43] V. R. Joseph, "Optimal ratio for data splitting," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 4, pp. 531–538, 2022, doi: 10.1002/sam.11583.
[44] M. Thulin, "Modern statistics with R: From wrangling and exploring data to inference and predictive modelling," *Modern Statistics with R: From Wrangling and Exploring Data to Inference and Predictive Modelling*, pp. 1–474, 2024, doi: 10.1201/9781003401339.
[45] The International HapMap Consortium, "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, no. 7164, pp. 851–861, 2007, doi: 10.1038/nature06258.
[46] V. Moreno, J. R. Gonzalez, and D. Pelegri, "SNPassoc: SNPs-Based Whole Genome Association Studies [R package]." [Online]. Available: https://cran.r-project.org/web/packages/SNPassoc/refman/SNPassoc.html. (Accessed: Oct. 31, 2025).
[47] E. Uffelmann *et al.*, "Genome-wide association studies," *Nature Reviews Methods Primers*, vol. 1, no. 1, p. 59, 2021, doi: 10.1038/s43586-021-00056-9.
[48] W. Lau, A. Ali, H. Maude, T. Andrew, D. M. Swallow, and N. Maniatis, "The hazards of genotype imputation when mapping disease susceptibility variants," *Genome Biology*, vol. 25, no. 1, pp. 1-17, 2024, doi: 10.1186/s13059-023-03140-3.
[49] E. König, J. S. Mitchell, M. Filosi, and C. Fuchsberger, "Impact of the inaccessible genome on genotype imputation and genome-wide association studies," *Human Molecular Genetics*, vol. 33, no. 14, pp. 1207–1214, Jul. 2024, doi: 10.1093/hmg/ddae062.
[50] L. Phan, "SNPs Classification and Terminology: dbSNP Reference SNP (rs) Gene and Consequence Annotation," in *Single Nucleotide Polymorphisms: Human Variation and a Coming Revolution in Biology and Medicine*, Eds., Cham: Springer International Publishing, 2022, pp. 3–12, doi: 10.1007/978-3-031-05616-1_1.
[51] B. S. Ko, S. B. Lee, and T. K. Kim, "A brief guide to analyzing expression quantitative trait loci," *Molecules and Cells*, vol. 47, no. 11, pp. 1-9, 2024, doi: 10.1016/j.mocell.2024.100139.
[52] P. J. Greer *et al.*, "A reassessment of Hardy-Weinberg equilibrium filtering in large sample Genomic studies," *medRxiv*, 2024, doi: 10.1101/2024.02.07.24301951.

[53] W. S. Pearman, L. Urban, and A. Alexander, "Commonly used Hardy – Weinberg equilibrium filtering schemes impact population structure inferences using RADseq data," *Molecular Ecology Resources*, vol. 22, no. 7, pp. 2599–2613, 2022, doi: 10.1111/1755-0998.13646.
[54] W. Chen *et al.*, "Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors," *Nature Communications*, vol. 12, no. 1, pp. 1–10, 2021, doi: 10.1038/s41467-021-27438-7.
[55] S. Wani and N. Alonso, "Quality control (QC) protocol for Genome Wide Association Study (GWAS) data," *HubLE Methods*, vol. 11, pp. 10–12, 2020, doi: 10.13140/RG.2.2.17494.88645.
[56] B. O. Petrazzini, H. Naya, F. Lopez-Bello, G. Vazquez, and L. Spangenberg, "Evaluation of different approaches for missing data imputation on features associated to genomic data," *BioData Min*, vol. 14, no. 1, pp. 1–13, 2021, doi: 10.1186/s13040-021-00274-7.
[57] C. C. et al. Chang, "PLINK 2.0 documentation—Population stratification (PCA): the randomized algorithm always mean-imputes missing genotype calls." [Online]. Available: https://www.cog-genomics.org/plink/2.0/strat. (Accessed: Sep. 30, 2025).
[58] T. Naito and Y. Okada, "Genotype imputation methods for whole and complex genomic regions utilizing deep learning technology," *Journal of Human Genetics*, vol. 69, no. 10, pp. 481–486, 2024, doi: 10.1038/s10038-023-01213-6.
[59] P. N. Basuki, J. P. S. Yulianto, and A. Setiawan, "Supplementary materials and datasets for 'Improving Genomic Classification via Pearson-Based SNP Selection: A Comparison of k-NN, SVM, and Random Forest," *Zenodo*, 2025, doi: 10.5281/zenodo.17271719.

BIOGRAPHIES OF AUTHORS

Prihanto Ngesti Basuki hold a bachelor's degree in the Faculty of Mathematics and Natural Sciences at Gadjah Mada University in 1992. He also received his Master's degree in Computer Science at Gadjah Mada University in 2004. He is currently studying for a Doctorate in Computer Science, Faculty of Information Technology, Satya Wacana Christian University, Indonesia. His research interests are in single nucleotide polymorphism, data analysis, python programming, and database programming. He can be contacted at email: ngesti@uksw.edu.

Sri Yulianto Joko Prasetyo earned his doctorate in Computer Science from Gadjah Mada University in 2013. Since 2008, he has been actively researching spatial data processing and remote sensing, with publications in Scopus-indexed international journals (SJR 0.8). His main interests include geospatial computing, with expertise in machine learning, remote sensing, spatial modeling, and software engineering. He is also the founder of Qua-edutechno, a technology-based startup focused on higher education quality management. He can be contacted at email: sri.yulianto@uksw.edu.

Adi Setiawan earned a Ph.D. in Statistics from Vrije Universiteit Amsterdam in 2007 and a Master of Mathematics from the same university in 1997. He is a Full Professor of Mathematics and a faculty member of the Master of Data Science program at Satya Wacana Christian University, Indonesia. His research focuses on bootstrapping, Bayesian statistics, survival analysis, SNP, and data analysis using R. He also works on data analysis in GIS, genetics, economics, and business. Currently, he serves as the head of the Faculty Quality Assurance Unit at Satya Wacana Christian University. He can be contacted at email: adi.setiawan@uksw.edu.