http://communications.science.ankara.edu.tr

Commun.Fac.Sci.Univ.Ank.Ser. A1 Math. Stat.
Volume 74, Number 3, Pages 492–502 (2025)
https://doi.org/10.31801/cfsuasmas.1517305
ISSN 1303-5991 E-ISSN 2618-6470

Research Article; Received: July 17, 2024; Accepted: March 19, 2025

A copula-based classification using agglomerated feature selection_extraction: an application in cervical cancer diagnostic

Necla KOÇHAN[1] and Ayyub SHEIKHI[2]

[1]Department of Mathematics, Faculty of Arts and Sciences, Izmir University of Economics, Izmir, TÜRKİYE
[2]Department of Statistics, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, IRAN

Abstract. The use of gene-expression datasets has significantly enhanced our understanding of complex diseases such as cancer. The importance of the relationship between genes in analyzing such datasets has been highlighted, indicating their crucial role in diagnosing the disease accurately. In this study, we investigate the associated copulas between attributes to extract fundamental block-related components. Subsequently, we perform a classification algorithm based on these components to classify a labeled target variable. Specifically, examining the practical implications and effectiveness of our approach in real-world scenarios, we provide a novel illustrative application in cervical cancer classification.

2020 Mathematics Subject Classification. 62H20, 62H05, 62H30.
Keywords. Copula, classification, feature selection, feature extraction, association measures, gene-expression.

1. Introduction

Classifying samples based on their gene-expression levels involves assigning a label or category to them using their gene-expression levels, which can be measured using high-throughput technologies such as microarrays or RNA-Sequencing (RNA-Seq). Unlike microarrays, which only measure the relative expression levels of a pre-determined set of genes, RNA-Seq allows researchers to sequence the entire transcriptome (all of the RNA molecules) in a sample and to measure the expression levels of all genes [32, 34]. This makes RNA-Seq a superior technique for gene-expression analysis as it is more comprehensive, accurate, and cost-efficient compared to microarray techniques. RNA-Seq is also more sensitive than microarrays, making it possible to detect low levels of expression that might be missed by microarray technologies.

Microarray technologies generate continuous data by measuring the fluorescence of DNA probes that correspond to specific genes, and these data can be used to determine the relative expression levels of the genes on the array. RNA-Seq, on the other hand, is a sequencing technique that is used to study gene-expression by determining the sequence of RNA molecules in a sample. The data generated by RNA-Seq are discrete and consist of counts of the number of times a particular RNA molecule was sequenced [28]. The fact that RNA-Seq data are discrete can have important implications for the statistical methods that are used to analyze the data. Some statistical methods, such as linear models, are designed to work with continuous data and may not be appropriate for analyzing RNA-Seq data. Other methods, such as generalized linear models, are specifically designed to handle discrete data and may be more suitable for analyzing RNA-Seq data [41, 48].

The classification of RNA-Seq data is a common task in the analysis of transcriptomic data and can be used to classify samples based on their gene-expression profiles and predict the functional roles of genes.

[1] necla.kochan@ieu.edu.tr -Corresponding author; 0000-0003-2355-4826.
[2] sheikhy.a@uk.ac.ir; 0000-0002-3731-6012.

Since RNA-Seq has a discrete nature, one approach to classify samples is to use discrete distributions such as Poisson or Negative Binomial distribution. For instance, [8] developed a new model called Negative Binomial Linear Discriminant Analysis (NBLDA). NBLDA is a variant of linear discriminant analysis (LDA) that is specifically designed to handle count data, such as the RNA-Seq data obtained from sequencing experiments. [49] introduced voomDDA classifiers, which extend the nearest shrunken centroids (NSC) and diagonal discriminant classifiers by incorporating variance modeling at the observational level (voom). The authors integrated the mean-variance relationship into these models using voom's precision weights. With the increasing popularity of machine-learning algorithms across various fields, researchers have focused on adopting these methods for RNA-Seq data classification tasks. For instance, [12] developed an R package called `MLSeq` including more than 80 machine-learning algorithms for gene expression data classification. For cutting-edge methods in RNA-seq data classification leveraging machine-learning techniques, the readers are referred to see the study by Jabeen et al. [18].

However, many studies related to RNA-Seq data classification assume that genes are independent and try to find a linear combination of genes that can best distinguish between different classes [45, 48]. Conversely, only a limited number of studies have accounted for the dependency structure or high collinearity among genes. For instance, [44] introduced a sparse version of Quadratic Discriminant Analysis (SQDA) method, which integrates regularized estimates of covariance matrices assumed to vary across different classes into the model. Zhang developed a method for classifying RNA-Seq data using Gaussian copulas, which are a type of statistical model that can be used to describe the dependence between variables [50]. The authors proposed using Gaussian copulas to model the dependencies between the expression levels of different genes to classify RNA-Seq data into different categories, such as different types of diseases. They evaluate the performance of their method on several RNA-Seq data sets and show that it is effective at accurately classifying the data. Others proposed different approaches to capture the dependencies between the expression levels of the gene can be found in these studies [23, 24]. They demonstrate the effectiveness of their method on several RNA-Seq data sets and show that it performs well in terms of classification accuracy. However, the nonlinear relationship between the genes, which could affect classification performance, has not been taken into account in any of these studies, which have only considered the linear relationship between the genes [11, 15, 29, 43]. Therefore, in this study, we incorporated copula functions in the classification algorithm for gene-expression data to investigate the classification performances, particularly when the relationship between genes is assumed to be nonlinear.

2. Materials and Methods

2.1. Feature Selection.

Feature selection is an essential step that must be completed before the classification phase as it can significantly impact the performance of the classification. Selecting the right features can help to improve the accuracy and efficiency of the classifier, by reducing the dimensionality of the data and eliminating irrelevant or redundant features [1, 3, 17]. In the context of RNA-Seq data, we followed the same strategy explained in [24]. This strategy is implemented using the `edgeR` package. The steps of this strategy are as follows:

(1) Filter out genes with low expression across all samples;
(2) Apply likelihood ratio test (LRT) on each remaining gene to test that they are differentially expressed between classes;
(3) Sort the genes according to their LRT statistic;
(4) Select the top n differentially expressed genes.

2.2. Copula Functions.

Several studies have shown that using copula functions significantly improves the performance of classification algorithms as they can account for inter-variable dependencies. For example, [46] explored a supervised classification method employing a hierarchical copula-based approach. Similarly, [31] conducted nonlinear random forest (RF) classification using a copula-based approach. Furthermore, [7] employed copula-based joint statistical models for image classification. [21] investigated the use of the maximal information coefficient as a classification index in a copula-based decision tree for two random variables. For more information on the application/use of copula functions in classification algorithms, additional resources include the works of [13, 19, 20, 51]. In this study, we employed copula functions where the genes are assumed to have a nonlinear dependency.

Now, assume that the random vector $\boldsymbol{X} = (X_1, X_2, ..., X_d)$ follows the joint multivariate distribution function $F_{X_1,X_2,...X_d} : \mathbb{R}^d \to [0,1]$ and $F_i : \mathbb{R} \to [0,1]$, $i = 1,2,...,d$, are the related marginal distribution function of X_i, $i = 1,2,...,d$. Based on the Sklar's theorem [42], there exists a grounded, uniformly marginal and d-increasing function

$$C : [0,1]^d \to [0,1]\,,$$

such that

$$F_{X_1,X_2,...X_d}(x_1, x_2, ..., x_d) = C(F_1(x_1), F_2(x_2), ..., F_d(x_d)). \tag{1}$$

As a matter of fact, the function $C : [0,1]^d \to [0,1]$ is called a copula of the random vector $\boldsymbol{X}$ whenever it joins the multivariate distribution function $F_{X_1,X_2,...X_d}$ to its marginals $F_i, i = 1,2,...,d$.

From (1), see also [9, 35], it follows, if the joint distribution function $F_{X1,..,Xd}$ is absolutely continuous, then one can obtain the density h of $(X_1, .., X_d)$ as

$$h\left(x_1, \ldots, x_d\right) = c\left(F_1\left(x_1\right), \ldots, F_d\left(x_d\right)\right) \prod_{k=1}^{d} f_k\left(x_k,\right) \tag{2}$$

where c is the density of the copula C and $f_k, k = 1, .., d$, are the respective densities of random variables $X_1, .., X_d$.

A copula function yields several association measures, among them Kendall's τ and Spearman's ρ which are defined respectively as

$$\tau_{X,Y} = 4\int_0^1 \int_0^1 C(u,v)dC(u,v) - 1.$$

and

$$\rho_{X,Y} = 12\int_0^1 \int_0^1 C(u,v)dudv - 3.$$

In the sequel, for the sake of the notational simplicity, we adopt the notations τ and ρ for $\tau_{X,Y}$ and $\rho_{X,Y}$ respectively. Applications of copulas and dependence measures in data science have been extensively discussed in the literature. See [2, 26] for a robust copula-based feature selection; [5, 6, 36] for copula-based clustering approaches. Also, [16] and [22] used copula functions in dimension reduction algorithms.

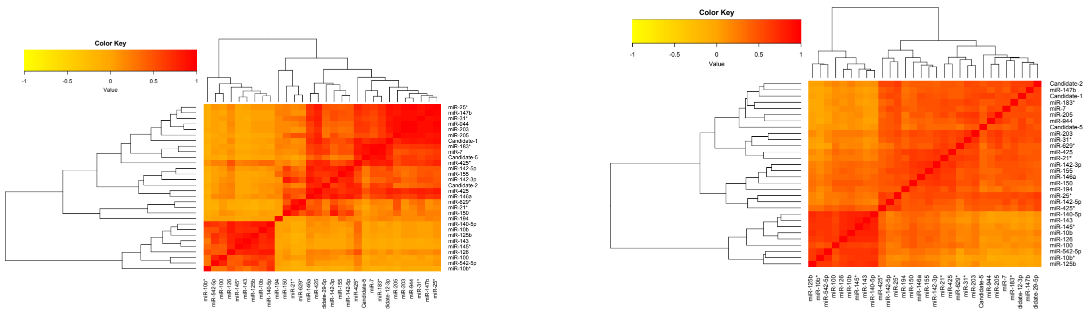
Especially in classification procedures, one may refer to [39] for using Gaussian copula in a supervised classification. Also in this context, [4] has introduced a copula-based classification for continuous and discrete data. [25] proposed a class-wise copula-based ensembling method for solving the multi-class segmentation problem. [31] investigated a nonlinear random forest classification using a copula-based approach. See also [10, 14, 38] for more applications of copulas in classification.

2.3. **Cervical Cancer Data.**

Cervical cancer data is one of the frequently used and publicly accessible RNA-seq datasets which has been provided for discovery of small RNA molecules (i.e., potential biomarkers) using cervical tumors and matched controls [47]. It comprises two categories: cancer and non-cancer, each containing 29 samples. Each sample contains counts for 714 distinct microRNAs obtained using RNA-seq.

3. Numerical Results

The nonlinear relationships between random variables play a crucial role in our classification. These relations can be captured by copulas. As a motivation for using copula dependence measures in classification, especially in cervical cancer diagnosis, we observed that some of the genes are nonlinearly dependent, while their linear correlations are negligible. For instance, the Pearson's correlation between "miR-21* " and "miR-223" is $\hat{r} = 0.08$ while their $\hat{\tau}$ is 0.69. Also for the two genes "miR-194" and "miR-150" we have $\hat{r} = 0.02$ while their $\hat{\tau} = 0.55$). Figure 1 compares the heat map plot of the 50 most relevant genes. In this section, we introduced two copula-based classification algorithms to give an account of the nonlinear dependence of variables.



(a) Pearson's correlation

(b) Kendall's association

Figure 1. Heatmap plot of the first 30 relevant genes: (a) Pearson's correlation, (b) Kendall's association

3.1. Feature Selection_Extraction Classification.

Recent developments in the field of high-dimensional studies have led to a renewed interest in dimension reduction, which is one of the critical steps in classification algorithms. Therefore, we proposed a combination of feature selection and feature extraction to remove the redundant features and extract high-variance-explained components. Algorithm 1 presents a pseudo code of our proposed approach.

Algorithm 1 Classification using feature selection_extraction

```
Data: Data set: matrix of the explanatory genes=$\boldsymbol{X}$, the class attribute=$y$
Result: Classification results
Step 1 - FS:
Select more relevant genes based on LRT method [37]
Step 2 - PCA:
Obtain the associated copula matrix $\hat{\Gamma}$ using either $\hat{\rho}_p$, $\hat{\rho}_s$ or $\hat{\tau}$
while convergence do
    update $\hat{v}_1$ with (3)
    update $\hat{v}_2, \ldots, \hat{v}_p$ using (4)
end while
Obtain the transformed new features using $\boldsymbol{u_j} = \boldsymbol{X\hat{v}_j}$ and construct $\boldsymbol{U} = (\boldsymbol{u_1 u_2 \ldots u_k})$
Step 3 - Classification:
Perform classification algorithms: KNN, LR for classifying $y$ using $U$ as the explanatory matrix.
```

By running the first step of our proposed algorithm, we extracted the most important attributes for the classification task based on the LRT method which has been proposed by Robinson et al. [37]. We only worked with the selected genes for $n = 50, 100, 200$. Table 1 reports selected genes for $n = 50$.

Table 1. The first 50 genes selected based on LRT method

The first 50 most relevant genes selected
miR-21* , miR-205, miR-7, miR-944, miR-155,, miR-10b*, miR-203, miR-143, Candidate-12-3p, miR-25*, miR-147b, miR-183*, miR-142-5p, miR-425, miR-31* Candidate-5, miR-125b, miR-146a, miR-10b, miR-145*, miR-142-3p, miR-425* miR-194 , miR-126, miR-629*, miR-542-5p, miR-100, miR-150, miR-210, Candidate-51-5p, miR-133b, miR-125a-5p, miR-223, miR-224, miR-204, miR-30c-1*, miR-31, miR-424, miR-222*, miR-195*, miR-130b, miR-497*, miR-21, miR-484, miR-200a*, miR-125b-1*, miR-92a-1*, miR-450a

The second step of the Algorithm 1 carries out two different principal component approaches to extract the components' set. In this regard, using a traditional correlation-based PCA as well as an association-based PCA we convey this step. For the correlation-based PCA, we used the traditional PCA approach considering the pairwise Pearson correlation matrix; while for the association-based, we followed the algorithm of Sheikhi et al. [40] by solving the following objective function

$$\hat{\boldsymbol{v}}_1 = \underset{\boldsymbol{v}\in\mathbb{R}^p,\|\boldsymbol{v}\|_2=1}{\text{argmax}} \ \boldsymbol{v}\hat{\boldsymbol{\Gamma}}\boldsymbol{v} \tag{3}$$

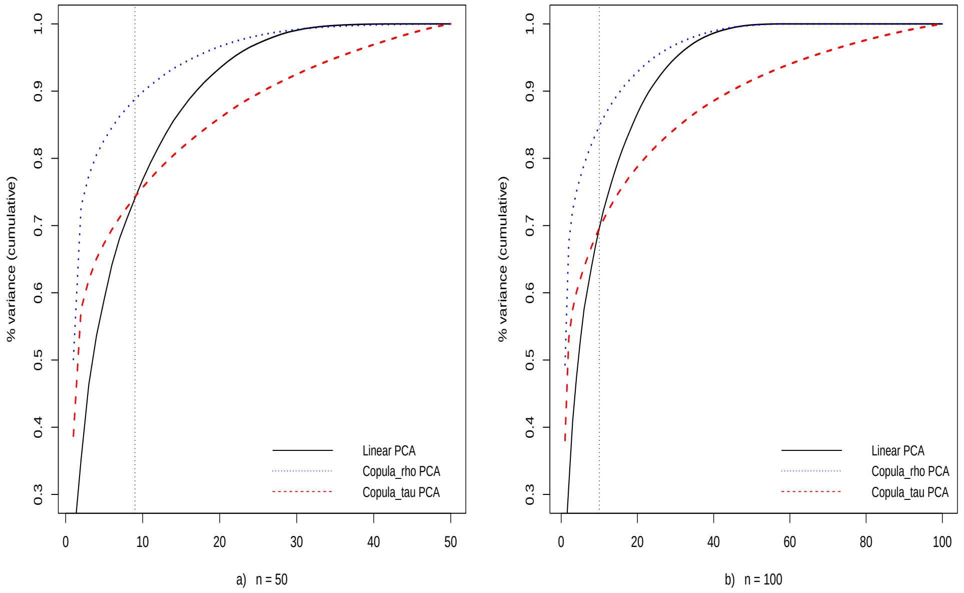
for the first principal component, and

$$\hat{\boldsymbol{v}}_k = \underset{\boldsymbol{v}\in\mathbb{R}^p,\|\boldsymbol{v}\|_2=1,\boldsymbol{v}^\top\boldsymbol{v}_j=0,\ j=1,\ldots,k-1}{\text{argmax}} \ \boldsymbol{v}\hat{\boldsymbol{\Gamma}}\boldsymbol{v} \quad k=2,3,...,p. \tag{4}$$

for the k-th principal component, $k = 2$, ..., p, where $\hat{\boldsymbol{\Gamma}}$, is either Spearman's or Kendall's association pairwise matrix. This procedure was carried out for $n = 50$ and 100 selected features. Additionally, three cases of the pairwise matrix: consisting of Pearson's $\hat{\rho}_p$, Spearman's $\hat{\rho}_s$ and Kendall's $\hat{\tau}$ were considered for the downstream analyses. For each of these matrices, we calculated the percentages of explained variances for some relevant number of extracted components. The results are summarized in Table 2. In the case of $n = 50$, the cumulative percentage of variance explained when we extracted 1,2,5,10,20,30,40, and 50 components that can be found in the left panel of Table 2. See also the right panel of Table 2, for extracting 1,5,10,20,40,60, 80, and 100 components when n=100. It can be seen from the table that the best performance belongs to Spearman's $\hat{\rho}_s$, but there exists a competition between Kendall's $\hat{\tau}$ and Pearson's $\hat{\rho}_p$ based on the number of the extracted components. More specifically, as Figure 2 reveals, in both $n = 50,\ 100$ cases, Spearman's $\hat{\rho}_s$ outperforms in contrary to those two Kendall's $\hat{\tau}$ and Pearson's $\hat{\rho}_p$. However, for $n = 50$, Kendall's $\hat{\tau}$ is better than Pearson's $\hat{\rho}_p$ if we extract up to 9 components (Figure 1a); while in the case of $n = 100$, for at most 10 extracted components, the explained variance form Kendall's $\hat{\tau}$ is better than Pearson's $\hat{\rho}_p$ (Figure 1b).

Table 2. Cumulative percentage of variance explained based on the number of extracted components: Left panel: $n = 50$, Right panel: n=100

$n = 50$				$n = 100$			
Components	$\hat{\rho}_p$	$\hat{\rho}_s$	$\hat{\tau}$	Components	$\hat{\rho}_p$	$\hat{\rho}_s$	$\hat{\tau}$
1	0.225	0.501	0.387	1	0.215	0.493	0.380
2	0.352	0.728	0.576	5	0.528	0.772	0.620
5	0.590	0.827	0.674	10	0.696	0.848	0.696
10	0.768	0.899	0.757	20	0.867	0.928	0.787
20	0.934	0.966	0.860	40	0.986	0.989	0.886
30	0.990	0.992	0.925	60	1.000	1.000	0.940
40	1.000	0.999	0.969	80	1.000	1.000	0.976
50	1.000	1.000	1.000	100	1.000	1.000	1.000



a) n = 50

b) n = 100

FIGURE 2. Percent of explained variance using Kendall's $\hat{\tau}$, Spearman's $\hat{\rho}_s$ and Pearson's $\hat{\rho}_p$: (a) $n = 50$, (b) $n = 100$

While the above estimators Pearsons $\hat{\rho}_p$, Spearman's $\hat{\rho}_s$ and Kendall's $\hat{\tau}$ serve as empirical measures of the relationships among all pairwise of these 50 genes, knowing the best-fitted copula would be beneficial to understand the type of relationship between each pair. To this, the empirical copula come into the role, in which for a sample $\{X_i, Y_i\}_{i=1}^n$ is given by

$$\hat{C}_n(u, v) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1} \left(\frac{R_i}{n+1} \leq u, \frac{S_i}{n+1} \leq v \right) \tag{5}$$

where R_i and S_i are respectively the ranks of X_i and Y_i in the sample, $\mathbf{1}(\cdot)$ is the indicator function, which equals 1 if the condition inside is true and 0 otherwise and $u, v \in [0, 1]$ are points in the copula space .

Applying the `VineCopula` R-package enables us to identify the most suitable copula for each pair. To prevent the presentation of a 50 x 50 matrix, Table 3 summarizes the best-fitted copulas along with their associated estimated Kendall's τ values for all pairwise comparisons of the first five genes. It is worth noting that the relationship between each pairwise gene can follow from copulas that capture the symmetric dependence such as Gaussian and t copula. Also, Table 3 reports some suitable copulas for capturing tail dependencies, including survival Joe and survival Gumbel.

TABLE 3. The best fitted copulas and their corresponding estimated Kendall's τ for all pairwise of the first five genes

	miR-21*	miR-205	miR-7	miR-944	miR-155
miR-21*	——	Survival Joe, (0.55)	Tawn type 1, (0.46)	Gaussian, (0.46)	Gaussian, (0.63)
miR-205	Survival Joe, (0.55)	——	Tawn type 2, (0.44)	Survival Gumbel, (0.63)	Gaussian, (0.42)
miR-7	Tawn type 1, (0.46)	Tawn type 2, (0.44)	——	Tawn type 1, (0.52)	Tawn type 1, (0.45)
miR-944	Gaussian, (0.46)	Survival Gumbel, (0.63)	Tawn type 1, (0.52)	——	t, (0.36)
miR-155	Gaussian, (0.63)	Gaussian, (0.42)	Tawn type 1, (0.45)	t, (0.36)	——

Finally, in the third step, we implemented the classification algorithms, i.e., K-nearest neighbor (KNN) and logistic regression (LR) to assess the performance of these classifiers using the explanatory matrix, $\boldsymbol{U}$.

3.2. **Agglomerated Feature Selection_Extraction Classification.** The main contribution of Algorithm 1 is its flexibility to regularize based on the desired information. In this subsection, we modify it with an agglomerative point of view. First of all, referring to Table 1 and Figure 2, by comparing the three correlation and concordance measures $\hat{\rho}_p$, $\hat{\rho}_s$, and $\hat{\tau}$, we observed that the Spearman's one, $\hat{\rho}_s$, had a better performance, so in this part, we only consider this measure. Some tuning parameters such as δ_1 and δ_2 may be implemented in a classification approach to regularize the running time, rate of convergence as well as accuracy of the classification.

In this manner, δ_1 can be considered as a tuning parameter in the feature selection step. We define this criterion using the Likelihood Ratio Statistic, LRS. Considering $\lambda_{(1)} \geq \lambda_{(2)} \geq \ldots \geq \lambda_{(n)}$ are respectively the LRS for the best relevant gene, the second best relevant gene,..., the last relevant (worst relevant) gene, then we define an agglomerated LRT feature selection ratio with the first k relevant genes as

$$Agg\_LRT = \sum_{i=1}^{k} \frac{\lambda_i}{\Lambda}, \tag{6}$$

where $\Lambda = \sum_{i=1}^{n} \lambda_i$.

More specifically, based on the number of variables, and number of cases, we can tune the value of δ_1 to acquire a level of accuracy in the feature selection step, see Algorithm 2 and Table 4. Similarly, δ_2 can be considered as the desired accuracy in the step of PCA_Classification. Having in mind that in a classification task, the accuracy which is the proportion of true results, either true positive or true negative can be obtained as follows: $a = \frac{TP+TN}{TP+TN+FP+FN}$, where TP, TN, FP, and FN are "true positive", "true negative", "false positive," and "false negative", respectively in a confusion matrix [40]. We use the agglomerative PCA_classification's accuracy as

$$Acc\_PCA = \sum_{i=1}^{k} \frac{a_i}{A}, \tag{7}$$

where $A = \sum_{i=1}^{n} a_i$ and $a_i, i = 1, \ldots,\ n$ is the accuracy of classification when we extract i principal components. As designed in Algorithm 2, we may proceed the feature selection and feature extraction steps to acquire the desired features and components based on the values of δ_1 and δ_2.

Algorithm 2 Agglomerated feature selection_extraction classification

```
Data: Data set: matrix of the explanatory genes=$\boldsymbol{X}$, the class attribute=$y$, threshold values $\delta_1$ and $\delta_2$.
Result:Number of first relevant genes in FS step and number of PCs in PCA step
Step 1 - FS:
Let $k = 1$;
while $Agg\_LRT \leq \delta_1$ do
    Perform $LRT$;
    Set $k = k + 1$
end while
Obtain the associated copula matrix $\hat{\boldsymbol{\Gamma}}$ using $\hat{\rho}_p$ for the selected relevant genes.
Step 2 - PCA Classification:
Update $\hat{\boldsymbol{v}}_1$ with 3;
while $Agg\_PCA \leq \delta_2$ do
    Set $p = p + 1$;
    Update $\hat{\boldsymbol{v}}_2, \ldots, \hat{\boldsymbol{v}}_p$ using 4;
    Obtain the transformed new features using $\boldsymbol{u}_j = \boldsymbol{X}\hat{\boldsymbol{v}}_j$ and construct $\boldsymbol{U} = (\boldsymbol{u}_1\boldsymbol{u}_2....\boldsymbol{u}_k)$
    Perform classification algorithms: KNN, LR and RF for classifying $y$ using $\boldsymbol{U}$ as the explanatory matrix.
end while
```

To better illustrate this approach, we used cervical cancer gene-expression data. Considering two values of δ_1 and δ_2, via Algorithm 2, we first identified the relevant genes which correspond to δ_1. We then applied the PCA_classification's step to obtain the most informative principal components. Table 4 presents a summary of the classification results. The table highlights that achieving a 95% accuracy in classification does not necessitate the use of the entire gene set or a large number of principal components. More specifically, the initial row of this table indicates that employing the first 198 most relevant genes (at least 75% for δ_1) necessitates only 13 principal components for achieving a classification accuracy of 95.56%. Also, in this row, KNN has the best performance among other classification methods. Similarly, the final row demonstrates that aiming for a 95% accuracy for both δ_1 and δ_2 requires the utilization of the top 400 relevant genes alongside 22 extracted principal components when random forest is implemented for classification.

Table 4. Number of top relevant genes in the feature selection step and number of principal components needed to achieve 95% of accuracy in classification

δ_1	k	δ_2	p	Best classification algorithm	Accuracy
75%	198	95%	13	KNN	95.56%
80%	231	95%	6	LR	96.37%
85%	272	95%	8	KNN	98.32%
90%	323	95%	16	LR	96.66%
95%	400	95%	22	RF	95.03%

4. Discussion and Conclusion

In this study, we investigated the impact of the nonlinear dependence between covariates, i.e. genes, in the classification problems using copula functions. More specifically, we explored the dependency between genes using copula functions, revealing essential block-related components to make a novel classification in cervical cancer diagnostics. We evaluated widely used classification algorithms based on these components, focusing on its efficacy in classifying cervical cancer patients. Our results showed that we can obtain high classification performance using a small number of genes and components. Furthermore, our approach overcomes the limitations of conventional linear methods by addressing nonlinear relationships among genes, enabling a deeper understanding of complex interactions within gene expression data. Therefore, we believe that our approach will contribute to the ongoing efforts to improve diagnostic and prognostic capabilities in the field of genomics, potentially paving the way for more precise and personalized treatment strategies.

The concept presented in this study has the potential for various extensions. Although we employed the LRT method to identify the most informative genes, it could be beneficial to investigate copula-based feature selection during the selection process [27]. It has been known that implementing the phenotype variables in such studies and their relationships with the genes will help to yield a better classification [33]. In this regard, investigating the nonlinear relationship between features in the set of phenotype variables, between features in the set of genotype variables, and between features in the set of phenotype variables and genes can be studied. The relationships among pairwise attributes in this study are derived from a combination of symmetric and asymmetric copulas. However, comparing the performance of different copulas, i.e., elliptical and Archimedean copulas, in such studies would be advantageous. Consequently, identifying a suitable dataset for this analysis would be of particular interest, especially within the context of medical datasets and healthcare studies. As another extension in our ongoing study, we aim to apply concordance measures to incorporate the nonlinear relationship into the classification model [30].

Author Contribution Statements All authors contributed equally and significantly in writing this article. All authors read and approved the final manuscript.

Declaration of Competing Interests All authors declare no conflicts of interest in this paper.

References

[1] Büyükkeçeci, M., Okur, M. C., A comprehensive review of feature selection and feature selection stability in machine learning, *Gazi Univ. J. Sci., 36(4)* (2023), 1506–1520, https://dx.doi.org/10.35378/gujs.993763.

[2] Chang, Y., Li, Y., Ding, A., Dy, J., A robust-equitable copula dependence measure for feature selection, In *Artificial Intelligence and Statistics* (2016), IEEE, pp. 84–92.

[3] Chen, R. C., Dewi, C., Huang, S. W., Caraka, R. E., Selecting critical features for data classification based on machine learning methods, *J. Big Data, 7(1)* (2020), 52, https://dx.doi.org/10.1186/s40537-020-00327-4.

[4] Chen, Y., A copula-based supervised learning classification for continuous and discrete data, *J. Data Sci., 14(4)* (2016), 769–782.

[5] Di Lascio, F. M. L., Coclust: An r package for copula-based cluster analysis, In *Recent Applications in Data Clustering*, BoD–Books on Demand, 2018, p. 93, https://dx.doi.org/10.5772/intechopen.74865.

[6] Di Lascio, F. M. L., Disegna, M., A copula-based clustering algorithm to analyse eu country diets, *Knowl.-Based Syst., 132* (2017), 72–84, https://dx.doi.org/10.1016/j.knosys.2017.06.004.

[7] Dong, H., Xu, X., Sui, H., Xu, F., Liu, J., Copula-based joint statistical model for polarimetric features and its application in polsar image classification, *IEEE Trans. Geosci. Remote Sens., 55(10)* (2017), 5777–5789, https://dx.doi.org/10.1109/TGRS.2017.2714169.

[8] Dong, K., Zhao, H., Tong, T., Wan, X., Nblda: Negative binomial linear discriminant analysis for rna-seq data, *BMC Bioinform., 17(1)* (2016), 369, https://dx.doi.org/10.1186/s12859-016-1208-1.

[9] Durante, F., Sempi, C., Principles of Copula Theory, CRC press, 2015.

[10] Elidan, G., Copula network classifiers (cncs), In *Artificial intelligence and statistics* (2012), PMLR, pp. 346–354.

[11] Fathi, H., AlSalman, H., Gumaei, A., Manhrawy, I. I. M., Hussien, A. G., El-Kafrawy, P., et al., An efficient cancer classification model using microarray and high-dimensional data, *Comput. Intell. Neurosci., 2021* (2021), https://dx.doi.org/10.1155/2021/7231126.

[12] Goksuluk, D., Zararsiz, G., Korkmaz, S., Eldem, V., Erturk Zararsiz, G., Ozcetin, E., Ozturk, A., Karaagaoglu, A. E., Mlseq: Machine learning interface for rna-sequencing data, *Comput. Methods Programs Biomed., 175* (2019), 223–231, https://dx.doi.org/10.1016/j.cmpb.2019.04.007.

[13] Hammami, N., Bedda, M., Nadir, F., Probabilistic classification based on copula for speech recognitation: an overview, In *2013 International Conference on Computer Applications Technology (ICCAT)* (2013), IEEE, pp. 1–3.

[14] Han, F., Zhao, T., Liu, H., Coda: High dimensional copula discriminant analysis, *J. Mach. Learn. Res., 14* (2013), 629–671.

[15] Hazra, S., Shaw, A. K., Das, P., Ghosh, A., Gene co expression analysis for identifying some regulatory genes in human lung cancer, In *2022 IEEE International Conference of Electron Devices Society Kolkata Chapter (EDKCON)* (2022), pp. 221–225, https://dx.doi.org/10.1109/EDKCON56221.2022.10032946.

[16] Houari, R., Bounceur, A., Kechadi, M.-T., Tari, A.-K., Euler, R., Dimensionality reduction in data mining: A copula approach, *Expert Syst. Appl., 64* (2016), 247–260, https://dx.doi.org/10.1016/j.eswa.2016.07.041.

[17] Hu, J., Pan, K., Song, Y., Wei, G., Shen, C., An improved feature selection method for classification on incomplete data: Non-negative latent factor-incorporated duplicate mic, *Expert Syst. Appl., 212* (2023), 118654, https://dx.doi.org/10.1016/j.eswa.2022.118654.

[18] Jabeen, A., Ahmad, N., Raza, K., Machine Learning-Based State-of-the-Art Methods for the Classification of RNA-Seq Data, Springer International Publishing, Cham, 2018, pp. 133–172, https://dx.doi.org/10.1007/978-3-319-65981-7-6.

[19] Jajuga, K., Papla, D., Copula functions in model based clustering, In *From Data and Information Analysis to Knowledge Engineering*, Springer, 2006, pp. 606–613.

[20] Karine, A., Toumi, A., Khenchaf, A., Hassouni, M. E., Multivariate copula statistical model and weighted sparse classification for radar image target recognition, *Comput. Electr. Eng., 84* (2020), 106633, https://dx.doi.org/10.1016/j.compeleceng.2020.106633.

[21] Khan, Y. A., Shan, Q. S., Liu, Q., Abbas, S. Z., A nonparametric copula-based decision tree for two random variables using mic as a classification index, *Soft Comput., 25(15)* (2021), 9677–9692,

https://dx.doi.org/10.1007/s00500-020-05399-1.

[22] Klüppelberg, C., Kuhn, G., Copula structure analysis, *J. R. Stat. Soc. Ser. B Stat. Methodol., 71(3)* (2009), 737–753, https://dx.doi.org/10.1111/j.1467-9868.2009.00707.x.

[23] Kochan, N., Tütüncü, G. Y., Giner, G., A new local covariance matrix estimation for the classification of gene expression profiles in high dimensional rna-seq data, *Expert Syst. Appl., 167* (2021), 114200, https://dx.doi.org/10.1016/j.eswa.2020.114200.

[24] Kochan, N., Tutuncu, G. Y., Smyth, G. K., Gandolfo, C. L., Giner, G., qtqda: quantile transformed quadratic discriminant analysis for high-dimensional rna-seq data, *PeerJ, 7* (2019), e8260, https://dx.doi.org/10.7717/peerj.8260.

[25] Kuiry, S., Das, N., Das, A., Nasipuri, M., Edc3: Ensemble of deep-classifiers using class-specific copula functions to improve semantic image segmentation, *arXiv preprint arXiv:2003.05710* (2020), https://dx.doi.org/10.48550/arXiv.2003.05710.

[26] Lall, S., Sinha, D., Ghosh, A., Sengupta, D., Bandyopadhyay, S., Stable feature selection using copula-based mutual information, *Pattern Recognit., 107* (2020), 107697, https://dx.doi.org/10.1016/j.patcog.2020.107697.

[27] Lall, S., Sinha, D., Ghosh, A., Sengupta, D., Bandyopadhyay, S., Stable feature selection using copula based mutual information, *Pattern Recognit., 112* (2021), 107697, https://dx.doi.org/10.1016/j.patcog.2020.107697.

[28] Law, C. W., Chen, Y., Shi, W., Smyth, G. K., voom: precision weights unlock linear model analysis tools for rna-seq read counts, *Genome Biol., 15(2)* (2014), R29, https://dx.doi.org/10.1186/gb-2014-15-2-r29.

[29] Lopes, M. B., Casimiro, S., Vinga, S., Twiner: correlation-based regularization for identifying common cancer gene signatures, *BMC Bioinform., 20(1)* (2019), 356, https://dx.doi.org/10.1186/s12859-019-2937-8.

[30] Mesiar, R., Kolesárová, A., Sheikhi, A., Convex concordance measures, *Fuzzy Sets Syst., 441* (2022), 366–377, https://dx.doi.org/10.1016/j.fss.2022.01.001.

[31] Mesiar, R., Sheikhi, A., Nonlinear random forest classification, a copula-based approach, *Appl. Sci., 11(15)* (2021), 7140, https://dx.doi.org/10.3390/app11157140.

[32] Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., Wold, B., Mapping and quantifying mammalian transcriptomes by rna-seq, *Nat. Methods, 5* (2008), 621–628, https://dx.doi.org/10.1038/nmeth.1226.

[33] Mowlaei, M. E., Shi, X., Fsf-ga: A feature selection framework for phenotype prediction using genetic algorithms, *Genes (Basel), 14(5)* (2023), 1059, https://dx.doi.org/10.3390/genes14051059.

[34] Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M., The transcriptional landscape of the yeast genome defined by rna sequencing., *Science, 320(5881)* (2008), 1344–1349, https://dx.doi.org/10.1126/science.1158441.

[35] Nelsen, R. B., An introduction to copulas, Springer Science & Business Media, 2006.

[36] Ozdemir, O., Allen, T. G., Choi, S., Wimalajeewa, T., Varshney, P. K., Copula-based classifier fusion under statistical dependence, *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(11)* (2017), 2740–2748, https://dx.doi.org/10.1109/TPAMI.2017.2774300.

[37] Robinson, M. D., McCarthy, D. J., Smyth, G. K., edger: a bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics, 26(1)* (2009), 139–140, https://dx.doi.org/10.1093/bioinformatics/btp616.

[38] Salinas-Gutiérrez, R., Hernández-Aguirre, A., Rivera-Meraz, M. J. J., Villa-Diharce, E. R., Supervised probabilistic classification based on gaussian copulas, In *Mexican International Conference on Artificial Intelligence* (2010), Springer, pp. 104–115.

[39] Salinas-Gutiérrez, R., Hernández-Aguirre, A., Rivera-Meraz, M. J. J., Villa-Diharce, E. R., Using gaussian copulas in supervised probabilistic classification, In *Soft Computing for Intelligent Control and Mobile Robotics*, Springer, 2010, pp. 355–372.

[40] Sheikhi, A., Mesiar, R., Holeňa, M., A dimension reduction in neural network using copula matrix, *Int. J. Gen. Syst., 51(5)* (2022), 1–16, https://dx.doi.org/10.1080/03081079.2022.2108029.

[41] Si, Y., Liu, P., Li, P., Brutnell, T., Model-based clustering of rna-seq data, *Bioinformatics, 30(2)* (2014), 197–205, https://dx.doi.org/10.1093/bioinformatics/btt632.

[42] Sklar, M., Fonctions de repartition an dimensions et leurs marges, *Publ. Inst. Stat. Univ. Paris, 8* (1959), 229–231.

[43] Sonmez, O. S., Dagtekin, M., Ensari, T., Gene expression data classification using genetic algorithm-based feature selection, *Turk. J. Electr. Eng. Comput. Sci., 29(7)* (2021), https://dx.doi.org/10.3906/elk-2102-110.

[44] Sun, J., Zhao, H., The application of sparse estimation of covariance matrix to quadratic discriminant analysis, *BMC Bioinform., 16(1)* (2015), 48, https://dx.doi.org/10.1186/s12859-014-0443-6.

[45] Tan, K. M., Petersen, A., Witten, D., Classification of RNA-seq Data, Springer International Publishing, Cham, 2014, pp. 219–246, https://dx.doi.org/10.1007/978-3-319-07212-8-11.

[46] Voisin, A., Krylov, V. A., Moser, G., Serpico, S. B., Zerubia, J., Supervised classification of multisensor and multiresolution remote sensing images with a hierarchical copula-based approach, *IEEE Trans. Geosci. Remote Sens., 52(6)* (2013), 3346–3358, https://dx.doi.org/10.1109/TGRS.2013.2272581.

[47] Witten, D., Tibshirani, R., Gu, S. G., et al., Ultra-high throughput sequencing-based small rna discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls, *BMC Biol., 8* (2010), 58, https://dx.doi.org/10.1186/1741-7007-8-58.

[48] Witten, D. M., Classification and clustering of sequencing data using a poisson model, *Ann. Appl. Stat., 5(4)* (2011), 2493–2518, https://dx.doi.org/10.1214/11-AOAS493.

[49] Zararsiz, G., Goksuluk, D., Klaus, B., Korkmaz, S., Eldem, V., Karabulut, E., Ozturk, A., voomdda: Discovery of diagnostic biomarkers and classification of rna-seq data, *PeerJ, 5* (2017), e3890, https://dx.doi.org/10.7717/peerj.3890.

[50] Zhang, Q., Classification of rna-seq data via gaussian copulas, *Stat, 6(1)* (2017), 171–183, https://dx.doi.org/10.1002/sta4.144.

[51] Zhang, Y., Wang, X., Liu, D., Li, C., Liu, Q., Cai, Y., Yi, Y., Yang, Z., Joint probability-based classifier based on vine copula method for land use classification of multispectral remote sensing data, *Earth Sci. Inform., 13(4)* (2020), 1079–1092, https://dx.doi.org/10.1007/s12145-020-00487-0.