Bitlis Eren Üniversitesi Fen Bilimleri Dergisi
BİTLİS EREN UNIVERSITY JOURNAL OF SCIENCE
ISSN: 2147-3129/e-ISSN: 2147-3188
VOLUME: 13 NO: 4 PAGE: 1335-1354 YEAR: 2024
DOI: 10.17798/bitlisfen.1561354

Intrusion Detection and Performance Analysis Using Copula Functions

Mehmet BURUKANLI[1], Musa ÇIBUK[2 *]

[1] *Bitlis Eren University, Rectorate, Department of Common Courses, Bitlis, Türkiye*
[2] *Bitlis Eren University, Department of Computer Engineering, Bitlis, Türkiye*
(ORCID: 0000-0003-4459-0455) (ORCID: 0000-0001-9028-2221)

Keywords: Intrusion Detection, Copula Functions, KDD'99 Dataset, Naive Bayes Classifier, mRMR.

Abstract

Nowadays, interest in technology is growing as technology advances and makes our jobs easier. These rapid technological advancements bring with them a slew of unwanted negative attacks, such as cyber-attacks and unauthorized access. To prevent such negative attacks, intrusion detection systems are frequently used. In this research, we make some suggestions for novel and reliable classifiers for intrusion detection systems that are based on copulas. Using copula-based classifiers, we hope to detect intrusion in computer networks. Student's-t, Gumbel, Clayton, Gaussian, Independent and Frank classifiers, which are frequently used in the literature, have been preferred as copula-based classifiers. These classifiers were used to perform classification on the KDD'99 dataset. The 10-fold cross-validation method has been used in the classification phase. When the experimental results were examined, the proposed Gaussian copula-based classifier outperformed state-of-the-art basic methods on the KDD'99 dataset with a success rate of 99.41%. As a direct consequence of this, classifiers based on the copula have shown promising results in the field of intrusion detection. Classifiers that are based on the copula have been found to be a competitive alternative to the most recent and cutting-edge fundamental approaches.

1. Introduction

Today, the internet is an important communication tool that provides the flow of information between both personal and business relationships. This communication tool has also brought security risks. In particular, e-commerce applications made over the internet are exposed to serious attacks. These attacks cause significant damage to companies by causing loss of workforce, time and product in critical business applications [1]. Computer viruses and malware are examples of a few of these attacks. As a result of the attacks, information is lost and information that should be kept confidential may be disclosed. Security vulnerabilities on the Internet can cause great harm to web-based companies and public services. For this reason, companies and public service institutions increase their security measures day by day and have to make larger investments in order to take precautions against new threats [1]–[3]. Therefore, the tools that ensure the security of computer systems are gaining more and more importance, and especially the importance of Intrusion Detection Systems (IDS) is increasing. IDS helps to protect information systems against all kinds of attacks made over the network and is called any software or hardware components that have warning characteristics [2], [4]. By using IDS, attacks made over the network can be detected and prevented by activating the relevant mechanisms. There are many methods for performing IDSs. Some of these methods can be listed as Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Tree (DT), and Ensemble Learning (EL). Apart from these methods, methods such as copula functions have also started to be applied [5]. In this study, attack detection has been performed by using copula functions, which is a new approach for the IDS domain. The classification algorithm inspired in this study has been first proposed by R. Salinas-Gutierrez et al. [6]. This classification algorithm, which is Gaussian copula-based, has been later generalized by M. Scavnicky [7] and a copula-based classification algorithm has been obtained. In this study, the use and performance of the

* Corresponding author: mcibuk@beu.edu.tr
Received: 04.10.2024, Accepted: 25.12.2024

algorithms developed on the basis of the copula-based classification approach, which has been generalized by M. Scavnicky [7], in the field of IDS have been examined. There have been many studies on intrusion detection in the literature. We can list some of these studies as follows. B. W. Masduki et al. [2], in their study, made attack detection using SVM. The success rate in R2L attacks was 96.08%. Ş. Sağıroğlu et al. [4] developed an ANN-based intelligent IDS in their study. The clever IDS they developed yielded very successful results. They tested the intelligent IDS using the KDD'99 dataset. They used 65536 samples from the KDD'99 dataset. The highest success rate they achieved was 97.92% and the lowest success rate was 81.93%. H. A. Sonawane et al. [8] proposed two methods in their study, namely Neural Networks (NN) and NN-based principal component analysis (PCA). The NN-based PCA method used a few features of the KDD'99 dataset, while the NN method used all the features of the KDD'99 dataset. By comparing these two methods, they observed that NN gave better results than PCA. The highest success rate of NN was 90.20%. M. Govindarajan et al. [9] performed attack detection using ensemble classifier (radial basis function (RBF)+SVM) in their study. The best success rate of RBF+DVM was 85.19%. A. Dastanpour et al. [10] performed attack detection using SVM, ANN and Genetic Algorithm (GA) algorithms in their study. When they applied the GA algorithm on SVM, they achieved 100% performance using 24 features of the KDD'99 dataset, while when they applied the GA algorithm on ANN, they achieved 100% performance on 18 features of the KDD'99 dataset. ANN+GA algorithm achieved better performance with fewer features. W. Wang et al. [11] performed attack detection using PCA in their study. The best success rate was 98.80%. S. Kumar et al. [12] used ANN to detect attacks in their study. The success rate on the KDD'99 dataset was 91.90%. They used 494021 samples for training and 311027 sample data for testing. J. Esmaily et al. [13] performed attack detection using DT and ANN in their study. In addition, DT and ANN algorithms were compared with each other. On the KDD'99 dataset; ANN gave better results with 99.71%. The success rate of the DT algorithm was 97.93%. Y. B. Bhavsar et al. [14] performed attack detection using DVM in their study. Normally, the success rate was 94.18%, while using 10-fold cross validation and RBF kernel, the success rate increased to 98.57%. G. Poojitha et al. [15] performed attack detection using ANN in their study. They used a total of 12723 samples from the KDD'99 dataset, 6363 samples for training and 6360 samples for testing. The success rate was 94.93%. B. Huyot et al. [5] performed online unsupervised attack detection on the Defense Advanced Research Projects Agency (DARPA) dataset by using copula theory or functions in their study. They achieved a success rate of 79% on the DARPA dataset. Although many [6], [7], [16]–[21] studies using copulas have been examined in the literature, no study on attack detection using copula functions has been found in the literature, except for this study. The contributions of this article have been given below.

- We propose novel and reliable classifiers (Student's-t, Gumbel, Clayton, Gaussian, Independent and Frank classifiers) for intrusion detection systems.
- We use the MrMR feature extraction technique for feature extraction.
- The proposed gaussian copula-based classifier outperformed state-of-the-art basic methods on the KDD'99 dataset with a success rate of 99.41%.
- We use the 10-fold cross-validation method to measure the performance of the proposed models.
- •We conclude that the copula-based classifier is a competitive alternative to several state-of-the-art methods.

2. Material and Method

2.1. Copula Functions

The term copula has been used in Latin to mean connection, relationship [7], [22]. The term copula was first proposed by Abe Sklar in 1959 [23]. Copulas have often been used to express (measure) the dependence between variables [5], [24]. The main purpose of copulas is to describe the interrelationship (dependency) of several random variables [25]. In addition, copulas are used to examine dependency structures among random variables and to obtain a multivariate distribution function [26]. Copula functions with their margins are used to construct the multivariate joint distribution function [6], [27]. In statistics science; copula are multivariate functions that provide the relationship between the common distribution functions of the random variable vector and the marginals of this distribution [22]. In other words; copulas are functions used to get a common distribution using marginal distributions [5], [22]. Copulas have been used in many application areas. At the beginning of these application areas are insurance [28], finance [29], statistics [30], economics [31], risk management [32] and security [3], [6], [33]. Copulas have an important place in applications because their elements in a class can be easily constructed, contain

large variables, and have good algebraic properties [34]. Copula functions have generally divided into two basic categories, elliptical and Archimedean copulas [35], [36].

$u = (u_1, u_2, u_3, \dots, u_n; \theta) \in [0,1]^n$ is the random variable vector and C is the joint distribution function. C has been shown in equation (1).

$$\mathrm{C}(\mathrm{u}_1, \mathrm{u}_2, \mathrm{u}_3, \dots, \mathrm{u}_\mathrm{n}; \theta) = \mathrm{P}(\mathrm{U}_1 \le \mathrm{u}_1, \mathrm{U}_2 \le \mathrm{u}_2, \mathrm{U}_3 \le \mathrm{u}_3, \dots, \mathrm{U}_\mathrm{n} \le \mathrm{u}_\mathrm{n}). \quad (1)$$

where θ is the copula parameter. Copulas can take a single parameter or more than one parameter according to their type. A two-dimensional (variable) copula function C, a continuous distribution function whose marginals are defined as $u = [0,1]$ and $C: [0,1]^2 \to [0,1]$ has the following properties.
$C(0,u) = C(u,0) = 0$ for $\forall_n \in [0,1]$
$C(1,u) = C(u,1) = u$ for $\forall_n \in [0,1]$
$C(v_1, v_2) - C(v_1, u_2) - C(u_1, v_2) + C(u_1, u_2) \ge 0$ for each $(u_1, u_2), (v_1, v_2) \in [0,1] * [0,1]$ with $u_1 \le v_1$ and $u_2 \le v_2$ [5], [37][22] [3].

On the other hand, the Sklar theorem is one of the most important theorems of copulas [37]. The Sklar theorem mentions the function of copulas in linking multivariate joint distribution functions with its (univariate) marginal distribution functions [22], [38]. This theorem has made copulas more understandable and has easily expressed the relationship between copulas and joint distribution functions [39], [40]. The Sklar theorem has been given in equation (2).

$$C(u_1, u_2, \dots, u_d) = F\left({F_1}^{-1}(u_1), {F_2}^{-1}(u_2), \dots, {F_d}^{-1}(u_d)\right) \quad (2)$$

where the marginal distributions of random variables $x_1, x_2, \dots, x_d$ have expressed as $u_1 = F_1(x_1), u_2 = F_2(x_2), \dots, u_d = F_d(x_d)$, respectively [41], [42]. Although there are many types of copula families; Gumbel, Independent, Clayton, Gaussian, Student's-t, and Frank copula families, which have been widely used in the literature, have been used in this study [18], [22], [43]. The copula density function has often been used to estimate the parameter of a copula [16]. Let F be given as the joint distribution function. Let f be the density function of the joint distribution function F. Let C, be the copula function. $F_1, F_2, F_3, \dots, F_n$ are the marginal distributions of the F joint distribution function. The copula density function c has been alculated as in equation (3) [23], [44].

$$c(u_1, u_2, \dots, u_n) = \frac{\partial^n C(u_1, u_2, \dots, u_n)}{\partial u_1, \partial u_2, \dots, \partial u_n} = \frac{f({F_1}^{-1}(u_1), {F_2}^{-1}(u_2), \dots, {F_n}^{-1}(u_n))}{\prod_{i=0}^{n} f_i({F_i}^{-1}(u_i))} . \quad (3)$$

The relationship between the multivariate density function $f(u_1, u_2, \dots, u_n)$ and the copula density function has been shown in equation (4) [37], [45].

$$f(u_1, u_2, \dots, u_n) = c(F_1(u_1), F_2(u_2), \dots, F_n(u_n); \alpha) \prod_{i=1}^{n} f_i(u_i) . \quad (4)$$

where f_i is the univariate density function of the marginal distribution function F_i. c is the copula density function. α is the copula parameter [3]. The definitions of copulas families used in this study are given below. Gaussian (Normal) copula is obtained from multivariate Gaussian distribution [37], [46], [47] [3]. Gaussian copula is radially symmetric in its dependence structure [21], [44], [48]. Gaussian copula cannot model tail dependence because it has upper tail dependence ($\lambda_u = 0$) and lower tail dependence ($\lambda_l = 0$) [21]. The general representation of Gaussian copula is expressed as $C_\rho^{Gaussian}$. Student's-t copula is obtained from multivariate Student's distributions [46], [47]. The general representation of Student's-t copula is expressed as $C_{v,\rho}^{Student's-t}$. Student's-t copula is radially symmetric in its dependence structure [21], [44], [48]. Student'-t copula models both lower tail and upper tail dependency [21]. The Archimedean copula family is one of the most important copula families. Archimedean copulas are frequently preferred because they are easy to construct, can model various dependency properties of copula families, and are easy to use in applications [48], [50]. One of the advantages that makes Archimedean copulas useful is that they have a closed form [51], [52]. Archimedean copula families are generated using the formula in equation (5) [44], [53].

$$C_\theta(u_1, u_2, \dots, u_n) = \varphi^{-1}\left(\sum_{i=1}^{n} \varphi(u_i)\right) \quad (5)$$

where θ is the dependency parameter and φ is the generator function. In this study, Archimedean copulas (Clayton, Gumbel, Frank and Independent) were used for intrusion detection [3]. Naive Bayes theorem is a method for calculating (overcome) data uncertainty [54], [55]. Naive bayes is a classification technique frequently used in machine learning and data mining [19], [56]. Naive bayes classifier is based on bayes rule and probability theorem [57]. This

classifier is a supervised learning method used to predict the class from the features of the dataset [58], [59]. Copula-based classifiers often make use of the naive bayes theorem. The naive bayes theorem has been given in equation (6). The expression given in equation (6) has been used as a classification tool.

$$P(S_d|X) = \frac{P(S_d)P(X|S_d)}{P(X)} \quad (6)$$

Maximum A Posterior (MAP) is an estimation method used in naive bayes theorem [60], [61]. The MAP classifier can be designed by comparing the posterior probability $P(S_d|X)$. In other words, the MAP can be constructed by looking for the class that maximizes the posterior probability [62]. Also, MAP is used to select the best probability [63]. Let $X = (X_1, X_2, , X_3, \dots, X_K)$ features of an object belong to class D. When classifying this object, X features is assigned to the class with the posterior highest probability in class D [6]. For continuous features; A Gaussian copula function can be used to model the dependency structure in the probability function. In this case; probability can be calculated as in equation (7) by using the Gaussian copula density in the expression given in equation (6).

$$P(S_d|X) = \frac{\Phi(F_1(X_1),F_2(X_2),\dots,F_K(X_K)|S_d)P(S_d)\prod_{k=1}^{K} f_k(X_k|S_d)}{f(X)} \quad (7)$$

where F_i, is the marginal distribution functions and f_i is the marginal density of the features. Φ denotes Gaussian copula density [6], [7]. It can be generalized as in equation (8) by writing any (Student's-t, Clayton, Frank, Gumbel, Independent) copula density instead of Gaussian copula density in equation (7) [6], [7].

$$P(S_d|X) = \frac{\mathrm{c}(F_1(X_1),F_2(X_2),\dots,F_K(X_K)|S_d)P(S_d)\prod_{k=1}^{K} f_k(X_k|S_d)}{f(X)} \quad (8)$$

where c represents any copula density. The expression given in equation (8) is expressed as a naive bayesian classifier. The naive bayesian classifier can be obtained using the Independent copula. Therefore, the copula-based classifier is its generalized version. Based on the expression given in equation (8), a copula-based MAP classifier can be constructed as in equation (9) [6], [7].

$$\widehat{S_D} = \arg\max_{S_d \in Y} \mathrm{c}(F_1(X_1), F_2(X_2), \dots, F_K(X_K)|S_d) \prod_{k=1}^{K} f_k(x_k|S_d)P(S_d) \quad (9)$$

While constructing the MAP classifier, if the Inference Functions for Margins (IFM) method has been applied; any (Gaussian, Student's-t, Clayton, Frank, Gumbel, Independent, etc.) copula, the marginals estimated during the application of the IFM method, their densities and previous experimental probabilities have been used equation (9) to obtain a MAP classifier. In the case of applying the *Canonical Maximum Likelihood* (CML) method; any copula has been obtained by substituting the experimental cumulative distribution functions, experimental densities, and previous experimental probabilities in equation (9) [7] [3].

2.2. KDD'99 (KDD Cup 1999) Dataset

The first study sponsored by DARPA was carried out by the Massachusetts Institute of Technology (MIT) Lincoln laboratory in 1998 [64], [65]. DARPA is a data set used to perform both training/learning and testing [64]. KDD'99 dataset has been obtained by preprocessing the DARPA dataset (feature extraction etc.) [66], [67]. KDD'99 has been widely used in the research of IDSs in recent years [66], [68]. The purpose of the widespread use of the KDD'99 dataset is to facilitate training and testing for intrusion detection [4], [69]. A lot of preprocessing has been needed before the DARPA dataset can be used for IDSs. In this study; the KDD'99 dataset, which has obtained by preprocessing the DARPA dataset, has been used. By using the KDD'99 dataset, training and test results can be obtained faster [69]. There are 38 attack types in total, 24 attack types in the KDD'99 training dataset and 14 attack types in the test dataset [65], [69]. KDD'99 is a dataset consisting of 41 features, 9 basic and 32 derived [65]. The KDD'99 dataset is divided into four categories: basic features, content features, time-based traffic features, and host-based traffic features [65], [70]. In this study, two different KDD'99 datafiles, KDD100 (kddcup.data.gz) and KDD10 (kddcup.data_10_percent.gz) have been used [71]. KDD100 consists of 4898431 samples and KDD10 consists of 494021 samples [71], [72]. The quantities and categories of normal and attack types in the KDD10 and KDD100 datasets have been shown in Table 1.

Table 1. Quantities and categories of normal and attack types found in the KDD10 and KDD100 datasets [12], [3]

	KDD10		KDD100	
Attack Type	**Quantity**	**Quantity**	**Quantity**	**Category**
apache2	-	-	-	-
back	2203	DoS	2203	DoS
buffer_overflow	30	U2R	30	U2R
ftp_write	8	R2L	8	R2L
guess_passwd	53	R2L	53	R2L
httptunnel	-	-	-	-
imap	12	R2L	12	R2L
ipsweep	1247	probe	12481	probe
land	21	DoS	21	DoS
loadmodule	9	U2R	9	U2R
mailbomb	-	-	-	-
mscan	-	-	-	-
multihop	7	R2L	7	R2L
named	-	-	-	-
neptune	107201	DoS	1072017	DoS
nmap	231	probe	2316	probe
normal	97278	normal	972781	normal
perl	3	U2R	3	U2R
phf	4	R2L	4	R2L
pod	264	DoS	264	DoS
portsweep	1040	probe	10413	probe
processtable	-	-	-	-
ps	-	-	-	-
rootkit	10	U2R	10	U2R
saint	-	-	-	-
satan	1589	probe	15892	probe
sendmail	-	-	-	-
smurf	280790	DoS	2807886	DoS
snmpgetattack	-	-	-	-
snmpguess	-	-	-	-
spy	2	R2L	2	R2L
sqlattack	-	-	-	-
teardrop	979	DoS	979	DoS
udpstorm	-	-	-	-
warezclient	1020	R2L	1020	R2L
warezmaster	20	R2L	20	R2L
worm	-	-	-	-
xlock	-	-	-	-
xsnoop	-	-	-	-
xterm	-	-	-	-
Total	**494021**		**4898431**	

On the other hand, the data amounts of KDD10 and KDD100 data sets and the percentages of normal and attack (DoS, Probe, U2R, R2L) types have been shown in Table 2.

Table 2. The data amounts of the data types in the KDD'99 dataset and the percentage ratios of normal and attack (U2R, R2L, DoS, Probe) types [71], [3]

Dataset Name	**Data Amount**	**Attack Types**				**Normal**
		U2R	**R2L**	**DoS**	**Probe**	
KDD10	494021	%0.01	%0.22	%79.23	%0.83	%19.69
KDD100	4898431	%0.001	%0.02	%79.27	%0.83	%19.85

2.3. Preprocessing Stages of Datasets

In this study, two datasets, KDD10 and KDD100, which are under the KDD'99 dataset, have been used. These datasets have consisted of labeled data. Therefore, it is also clear to which attacks the data in the KDD10 and KDD100 datasets belong. The “protocol_type”, “service” and “flag” fields in these data sets have been in text (string) format, while the other fields have been in numerical format. In order to be able to operate on the data sets in our study, all the fields in the data sets must be in numerical format. Therefore, numerical values have been given to each of the “protocol_type”, “service” and “flag” fields in text format. In the "attack_type" field in the KDD10 and KDD100 data sets, a numerical value of 1 has been given for "normal traffic", while a numerical value of 2 has been given to other "all attack types". The purpose of doing this is to determine whether it is an attack rather than the type of attack. While converting the "protocol_type" names in the KDD10 and KDD100 datasets to digital format, icmp, tcp and udp protocols have been digitized 1, 2 and 3 respectively. The conversion of the "flag" names in the KDD10 and KDD100 datasets to numeric format has been shown in Table 3.

Table 3. Converting "flag" names in KDD10 and KDD100 datasets to numeric format [3]

Flag	Numerical Value
OTH	1
REJ	2
RSTO	3
RSTOS0	4
RSTR	5
S0	6
S1	7
S2	8
S3	9
SF	10
SH	11

The conversion of "service" names in the KDD10 and KDD 100 datasets to numeric format has been shown in Table 4.

Table 4. Converting "service" names in KDD10 and KDD100 datasets to numeric format [3]

KDD10 dataset		KDD100 dataset	
Service Name	**Numerical Value**	**Service Name**	**Numerical Value**
IRC	1	IRC	1
X11	2	X11	2
Z39_50	3	Z39_50	3
auth	4	aol	4
bgp	5	auth	5
courier	6	bgp	6
csnet_ns	7	courier	7
ctf	8	csnet_ns	8
daytime	9	ctf	9
discard	10	daytime	10
domain	11	discard	11
domain_u	12	domain	12
echo	13	domain_u	13
eco_i	14	echo	14
ecr_i	15	eco_i	15
efs	16	ecr_i	16
exec	17	efs	17
finger	18	exec	18
ftp	19	finger	19
ftp_data	20	ftp	20
gopher	21	ftp_data	21
hostnames	22	gopher	22
http	23	harvest	23
http_443	24	hostnames	24
imap4	25	http	25
iso_tsap	26	http_2784	26
klogin	27	http_443	27
kshell	28	http_8001	28
ldap	29	imap4	29
link	30	iso_tsap	30
login	31	klogin	31
mtp	32	kshell	32
name	33	ldap	33
netbios_dgm	34	link	34
netbios_ns	35	login	35
netbios_ssn	36	mtp	36
netstat	37	name	37
nnsp	38	netbios_dgm	38
nntp	39	netbios_ns	39
ntp_u	40	netbios_ssn	40
other	41	netstat	41
pm_dump	42	nnsp	42
pop_2	43	nntp	43
pop_3	44	ntp_u	44
printer	45	other	45
private	46	pm_dump	46
red_i	47	pop_2	47
remote_job	48	pop_3	48
rje	49	printer	49
shell	50	private	50
smtp	51	red_i	51
sql_net	52	remote_job	52

Table 4 (Continuous). Converting "service" names in KDD10 and KDD100 datasets to numeric format [3]

ssh	53	rje	53
sunrpc	54	shell	54
supdup	55	smtp	55
systat	56	sql_net	56
telnet	57	ssh	57
tftp_u	58	sunrpc	58
tim_i	59	supdup	59
time	60	systat	60
urh_i	61	telnet	61
urp_i	62	tftp_u	62
uucp	63	tim_i	63
uucp_path	64	time	64
vmnet	65	urh_i	65
whois	66	urp_i	66
		uucp	67
		uucp_path	68
		vmnet	69
		whois	70

2.4. Application of mRMR Method to KDD10 Data Set

mRMR is a method for selecting the most relevant features among the features in a dataset and reducing redundancy [73], [74]. mRMR is an entropy (disorder) based feature selection method. Entropy calculates the uncertainty in a random feature. Entropy produces values between 0-1. As seen in Table 5, the mRMR method has been applied on the KDD10 dataset, and the features in the dataset have been ranked according to their importance. According to these listed features, classification has been done using copula functions.

Table 5. Ranking of features according to importance when mRMR_miq criterion is applied to KDD10 dataset [3]

Raw Order		**After implemented mRMR**	
Feature Name	**Feature No**	**Feature Name**	**mRMR_miq**
duration	1	count	23
protocol_type	2	dst_bytes	6
service	3	duration	1
flag	4	dst_host_count	32
src_bytes	5	src_bytes	5
dst_bytes	6	srv_count	24
land	7	dst_host_srv_count	33
wrong_fragment	8	flag	4
urgent	9	service	3
hot	10	protocol_type	2
num_failed_logins	11	land	7
logged_in	12	wrong_fragment	8
num_compromised	13	urgent	9
root_shell	14	hot	10
su_attempted	15	num_failed_logins	11
num_root	16	logged_in	12
num_file_creations	17	num_compromised	13
num_shells	18	root_shell	14
num_access_files	19	su_attempted	15
num_outbound_cmds	20	num_root	16
is_host_login	21	num_file_creations	17
is_guest_login	22	num_shells	18
count	23	num_access_files	19
srv_count	24	num_outbound_cmds	20
serror_rate	25	is_host_login	21
srv_serror_rate	26	is_guest_login	22
rerror_rate	27	serror_rate	25
srv_rerror_rate	28	srv_serror_rate	26
same_srv_rate	29	rerror_rate	27
diff_srv_rate	30	srv_rerror_rate	28
srv_diff_host_rate	31	same_srv_rate	29
dst_host_count	32	diff_srv_rate	30
dst_host_srv_count	33	srv_diff_host_rate	31

Table 5 (Continuous). Ranking of features according to importance when mRMR_miq criterion is applied to KDD10 dataset [3]

dst_host_same_srv_rate	34	dst_host_same_srv_rate	34
dst_host_diff_srv_rate	35	dst_host_diff_srv_rate	35
dst_host_same_src_port_rate	36	dst_host_same_src_port_rate	36
dst_host_srv_diff_host_rate	37	dst_host_srv_diff_host_rate	37
dst_host_serror_rate	38	dst_host_serror_rate	38
dst_host_srv_serror_rate	39	dst_host_srv_serror_rate	39
dst_host_rerror_rate	40	dst_host_rerror_rate	40
dst_host_srv_rerror_rate	41	dst_host_srv_rerror_rate	41

2.5. Feature Selection

As can be seen in Table 5, after the features have been ranked in order of importance using the mRMR_miq feature selection criterion on the KDD10 dataset, feature selection has been started based on the first (23rd) feature. The first feature has been taken and the relationship status of the other features has been examined according to this feature. For example, the 1st feature (23rd) has been chosen at first. Then, the accuracy rate in the classification process has been i by taking the 1st feature and the 2nd feature (6th). Accuracy rates have been obtained using this method when each subsequent feature has been added. This situation has continued until the last feature (feature 41). The classification process has been completed by obtaining the three best performance rates and the features used for each dataset. In Figure 1 has been shown the relationship between the total elapsed time and the calculated percent while selecting the feature on the KDD10 dataset.

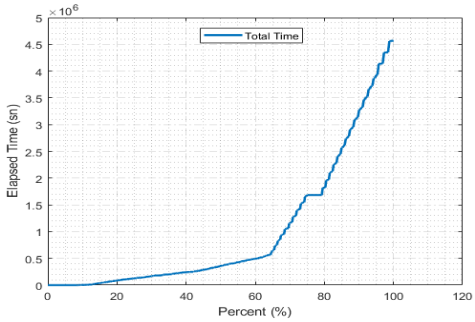

Figure 1. The relationship between the total calculated percentage depending on the total elapsed time while selecting the feature on the KDD10 dataset

As seen in Figure 1, the total time taken for all the features was while classifying on the KDD10 dataset has been calculated. During the calculation, as has stated above, a new feature has been added to the feature set each time, and the accuracy rates have been calculated. The amount of time required to calculate anything rises proportionally with the number of additional features provided. In particular, it has been observed that the calculated features spend much more time after the percentage rate is 80%. The calculation of the time elapsed between two features according to the use of IFM/CML methods with each copula family on the KDD10 dataset has been shown in Figure 2.

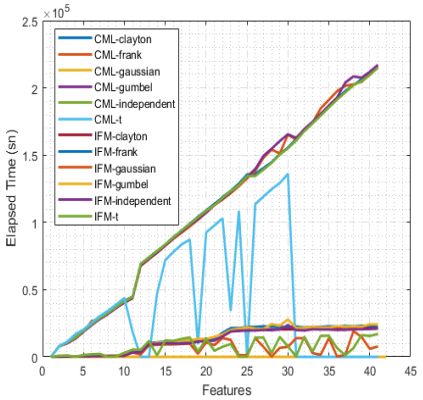

Figure 2. Calculation of the elapsed time between two features according to the use of IFM/CML methods with each copula family on the KDD10 dataset

As can be seen in Figure 2, the elapsed time between two features has been calculated at each step used while classifying the KDD10 dataset. While making the calculation, the accuracy rates have been calculated by adding the features separately as shown in Figure 2. The variation of features according to total elapsed time according to the use of IFM/CML methods with each copula family on the KDD10 data set has been shown in Figure 3.

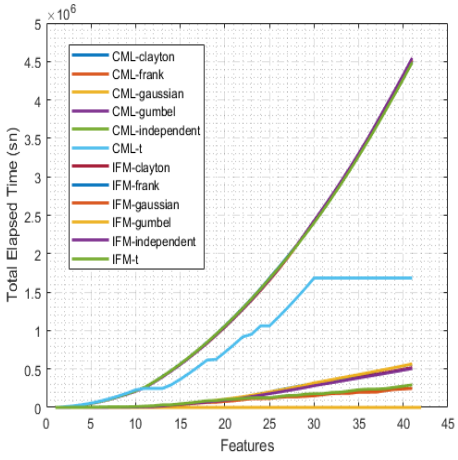

Figure 3. The variation of the characteristics according to the total elapsed time according to the use of IFM/CML methods with each copula family on the KDD10 dataset

As can be seen in Figure 3, while classifying on the KDD10 dataset, the total elapsed time for each copula family has been calculated by adding the time elapsed as a new feature has been added (to the previously calculated times). When all features have been used on the KDD10 dataset, the copula family that achieved the best performance has been the Gaussian copula using the IFM method. When all of the features were applied to the KDD10 dataset, the family of copulas known as the Gaussian copula utilizing the IFM approach produced the best results. The performance rates of the Gaussian copula family, which has shown the best performance for 41 features of the KDD10 dataset, according to the feature series ranked by mRMR have been shown in Figure 4.

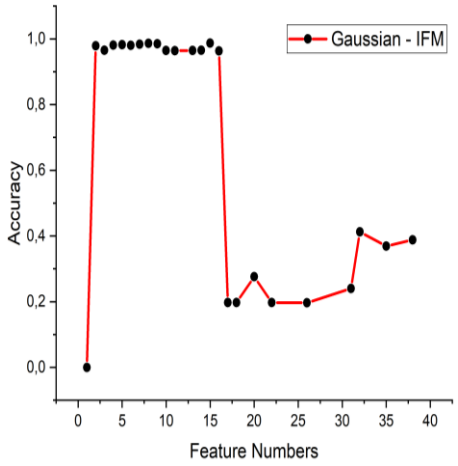

Figure 4. Performance ratios of the best performing Gaussian copula family for 41 features of the KDD10 dataset by feature series ranked by mRMR

As can be seen in Figure 4, the accuracy rates obtained depending on the number of Gaussian copula and IFM method features have been calculated on the KDD10 dataset. When taking into consideration the accuracy rates of the Gaussian copula, it has been shown that the best success rates have been reached between the second and the fifteenth features. This was discovered after taking into consideration the accuracy rates. This demonstrates to us that the first 15 features in the feature set, when applied with the IFM approach and the Gaussian copula, will produce relevant results when applied to the problem-solving process.

3. Results and Discussion

In this paper, intrusion detection has been made using copula-based classifiers and it has been investigated which copula has the best performance. The performance criterion used in this study has been given in equation (10) [3], [75].

$$Accuracy = \frac{True\ Positive\ (TP)+True\ Negative\ (TN)}{True\ Positive\ (TP)+False\ Negative\ (FN)+False\ Positive\ (FP)+True\ Negative\ (TN)} \quad (10)$$

In Figure 5 has been shown the application steps of copula-based classifiers to the KDD'99 dataset.

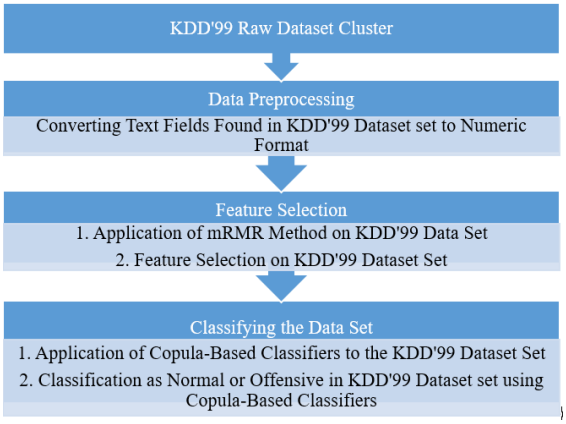

Figure 5. Application stages of copula-based classifiers to KDD'99 dataset

3.1. Application 1: KDD10

The KDD10 dataset was used in the first application because it contains less data but includes all samples. As a result, time was saved in feature selection. Gumbel, Independent, Clayton, Gaussian, Student's-t, and Frank copula families, as well as CML and IFM methods, were used to detect attacks in Application 1. 1%, 5%, 10%, and 50% of the KDD10 dataset were trained on a computer equipped with an Intel CPU (Xeon E5620), 16 GB RAM and Quadro K2000 GPU. Due to the large amount of data, 100% of the KDD10 dataset was trained on an HP-Z840 workstation with 2 10-cores Intel (Xeon E52687Wv3) processors, 64GB RAM and Quadro P5000 GPU. The trainings were held in the MATLAB environment. In the classification phase, the 10-fold cross-validation method was used. Using the confusion matrix found in Table 8 as the basis for the classifier evaluation metrics allowed for its acquisition. In the KDD10 dataset, certain percentages of each normal and attack type have been taken. 1% of the KDD10 dataset normally consists of 4940 data. But in this study, 4956 data have been used by taking 1% of each attack type. Purpose of this; it is to ensure that examples of all attack types are found while training the dataset. If the percentages of each attack type have been not taken at the same rate, low learning situations would occur in some attack types and excessive learning situations in others. This is true for 100%, 50%, 10% and 5% of the dataset. In Table 6, it has been demonstrated that the quantities of each attack type included in the KDD10 dataset correspond to the predetermined percentages (%) of the dataset.

Table 6. Quantities of each attack type found in the KDD10 dataset [3]

Attack Type	KDD10 (%100)	KDD10 (%50)	KDD10 (%10)	KDD10 (%5)	KDD10 (%1)
smurf	280790	140395	28079	14040	2808
neptune	107201	53601	10721	5361	1073
normal	97278	48639	9728	4864	973
back	2203	1102	221	111	23
satan	1589	795	159	80	16
ipsweep	1247	624	125	63	13
portsweep	1040	520	104	52	11
warezclient	1020	510	102	51	11
teardrop	979	490	98	49	10
pod	264	132	27	14	3
nmap	231	116	24	12	3
guess_passwd	53	27	6	3	1
buffer_overflow	30	15	3	2	1
land	21	11	3	2	1
warezmaster	20	10	2	1	1
imap	12	6	2	1	1
rootkit	10	5	1	1	1
loadmodule	9	5	1	1	1
ftp_write	8	4	1	1	1
multihop	7	4	1	1	1
phf	4	2	1	1	1
perl	3	2	1	1	1
spy	2	1	1	1	1
Total	**494021**	**247016**	**49411**	**24713**	**4956**

The amount of data has been calculated by taking the same percentages of each of the 23 attack types in the KDD10 dataset. For example; the data amount of the back attack type consists of 2203 samples in data set. Decimal numbers have been rounded. For example; While 1% of 2203, 22.03 samples should be taken from the back attack type, 23 samples have been taken due to rounding. The same is true for any other attack type. In Table 7, the success rates of the copula families with the best three performances have been shown by using 1% of KDD10 data set. Here, among the copulas that showed the same success, those who achieved this success at least have been considered more successful.

Table 7. Success rates of the best three performing copula families using the 1% rates of KDD10 dataset

Copula Family	Method	TP	TN	FP	FN	Accuracy (%)	Features Used
gumbel	**IFM**	**973**	**3920**	**63**	**0**	**98.73**	**“23 6 1 32 5 24 33 4 3 2”**
gumbel	IFM	973	3920	63	0	98.73	“23 6 1 32 5 24 33 4 3 2 7”
independent	IFM	973	3920	63	0	98.73	“23 6 1 32 5 24 33 4 3 2 7”
gaussian	IFM	973	3920	63	0	98.73	“23 6 1 32 5 24 33 4 3 2 10 12 22”
gaussian	IFM	973	3919	64	0	98.71	“23 6 1 32 5 24 33 4 3 2 10 12”
independent	IFM	973	3919	64	0	98.71	“23 6 1 32 5 24 33 4 3 2”
gaussian	IFM	973	3918	65	0	98.69	“23 6 1 32 5 24 33 4 3 2 10 12 22 29 31”
clayton	IFM	973	3918	65	0	98.69	“23 6 1 32 5 24 33 4 3 2 7”

As seen in Table 7, the best success rate has been obtained as 98.73% with Gumbel, Independent and Gaussian copulas using the IFM method. For the Gumbel copula family, this performance has been achieved using the “23 6 1 32 5 24 33 4 3 2” and “23 6 1 32 5 24 33 4 3 2 7” feature sets. Between these two feature sets, it should be preferred that shows the same performance with less features. While the Independent copula has achieved this success rate with the characteristics of “23 6 1 32 5 24 33 4 3 2 7”,

the Gaussian copula family has achieved this with the characteristics of "23 6 1 32 5 24 33 4 3 2 10 12 22". In this case, the Gumbel copula family has achieved the best performance by using fewer features than the Independent copula family and the Gaussian copula family. For 1% of the KDD10 dataset, Gumbel copula family, IFM method and "23 6 1 32 5 24 33 4 3 2" features should be preferred. In Table 8, the success rates of the copula families with the best three performances have been shown by using 5% of KDD10 dataset.

Table 8. The success rates of the best three performing copula families using 5% of KDD10 dataset

Copula Family	Method	TP	TN	FP	FN	Accuracy (%)	Features Used
gaussian	**IFM**	**4788**	**19707**	**142**	**76**	**99.12**	**"23 6"**
frank	**IFM**	**4788**	**19707**	**142**	**76**	**99.12**	**"23 6"**
independent	**IFM**	**4788**	**19707**	**142**	**76**	**99.12**	**"23 6"**
clayton	**IFM**	**4788**	**19707**	**142**	**76**	**99.12**	**"23 6"**
t	**IFM**	**4788**	**19707**	**142**	**76**	**99.12**	**"23 6"**
gumbel	**IFM**	**4788**	**19707**	**142**	**76**	**99.12**	**"23 6"**
independent	IFM	4853	19637	212	11	99.10	"23 6 1 32 5 24 33 4 3"
gumbel	IFM	4794	19692	157	70	99.08	"23 6 1"

As seen in Table 8, the best success rate with 99.12% has been obtained by using the IFM method with the characteristics of "23 6" for the Gaussian, Frank, Clayton, Gumbel, Independent and Student_t copula families. For 5% of the KDD10 data set, any of the Student_t, Gumbel, Clayton, Gaussian, Independent, and Frank copula families, IFM method and "23 6" features should be preferred. In Table 9, the success rates of the best three performing copula families have been shown by using 10% of KDD10 dataset.

Table 9. Success rates of the best three performing copula families using 10% of KDD10 dataset

Copula Family	Method	TP	TN	FP	FN	Accuracy (%)	Features Used
gaussian	**IFM**	**9585**	**39373**	**310**	**143**	**99.08**	**"23 6"**
frank	**IFM**	**9585**	**39373**	**310**	**143**	**99.08**	**"23 6"**
clayton	**IFM**	**9585**	**39373**	**310**	**143**	**99.08**	**"23 6"**
gumbel	**IFM**	**9585**	**39373**	**310**	**143**	**99.08**	**"23 6"**
independent	**IFM**	**9585**	**39373**	**310**	**143**	**99.08**	**"23 6"**
t	**IFM**	**9585**	**39373**	**310**	**143**	**99.08**	**"23 6"**
gaussian	IFM	9585	39373	310	143	99.08	"23 6 1"
frank	IFM	9585	39373	310	143	99.08	"23 6 1"
independent	IFM	9585	39373	310	143	99.08	"23 6 1"
clayton	IFM	9585	39373	310	143	99.08	"23 6 1"
gumbel	IFM	9585	39373	310	143	99.08	"23 6 1"
t	IFM	9584	39374	309	144	99.08	"23 6 1"
gumbel	IFM	9685	39260	423	43	99.06	"23 6 1 32 5 24 33 4 3"
independent	IFM	9617	39309	374	111	99.02	"23 6 1 32 5 24 33 4 3"

As seen in Table 9, the best success rate with 99.08% has been obtained by using the "23 6" and "23 6 1" feature sets for the Gumbel, Independent, Clayton, Gaussian, Student_t, and Frank copula families and the IFM method. For the Gaussian, Frank, Clayton, Gumbel, Independent and Student_t copula families, 23th and 6th features with a smaller number of features should be preferred. For 10% of the KDD10 dataset, any of the Gumbel, Independent, Clayton, Gaussian, Student_t, and Frank copula families, IFM method and "23 6" features should be preferred. In Table 10, the success rates of the best three performing copula families have been shown by using 50% of KDD10 dataset.

Table 10. Success rates of the best three performing copula families using 50% of KDD10 dataset

Copula Family	Method	TP	TN	FP	FN	Accuracy (%)	Features Used
gumbel	**CML**	**48174**	**196287**	**2090**	**465**	**98.97**	**"23 6 1 32 5 24 33 4 3 2 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 25 26 27 28 29 30"**
gumbel	CML	48250	196171	2206	389	98.95	"23 6 1 32 5 24 33 4 3 2 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 25 26 27 28"
gaussian	IFM	48128	196301	2076	511	98.95	"23 6 1 32 5 24 33 4 3 2 7 10 11 12"

As seen in Table 10, the Gumbel copula family has the best success rate with 98.97% by using "23 6 1 32 5 24 33 4 3 2 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 25 26 27 28 29 30" features and the CML method. The Gaussian copula family, on the other hand, has achieved a success rate of 98.95% using the IFM method with the characteristics of "23 6 1 32 5 24 33 4 3 2 7 10 11 12". For 50% of KDD10 dataset, Gumbel copula family, CML method and "23 6 1 32 5 24 33 4 3 2 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 25 26 27 28 29 30" features should be preferred. In Table 11, the success rates of the copula families with the best three performances have been shown by using the 100% rates of KDD10 dataset.

Table 11. Success rates of the best three performing copula families using 100% of KDD10 dataset

Copula Family	Method	TP	TN	FP	FN	Accuracy (%)	Features Used
gaussian	**IFM**	**96325**	**391193**	**5550**	**953**	**98.68**	**"23 6 1 32 5 24 33 4 3 2 7 9 10 11"**
gaussian	IFM	95710	391601	5142	1568	98.64	"23 6 1 32 5 24 33 4"
gaussian	IFM	96183	390323	6420	1095	98.48	"23 6 1 32 5 24 33 4 3"

As seen in Table 11, the Gaussian copula family has obtained the best success rate with 98.68% by using the "23 6 1 32 5 24 33 4 3 2 7 9 10 11" features and the IFM method. Gaussian copula family, IFM method and "23 6 1 32 5 24 33 4 3 2 7 9 10 11" features should be preferred for 100% of the KDD10 dataset.

3.2. Application 2: KDD100

In Application 2; Attack detection has been performed using Gaussian, Gumbel, Clayton, Student's-t, Independent and Frank copula families and CML and IFM methods. As has been seen in Table 12, the "23 6 1 32 5 24 33 4 3 2 7 9 10 11" feature set, the "23 6 1 32 5 24 33 4" feature set and the feature set "23 6 1 32 5 24 33 4 3" has been used in this study. These feature sets in the KDD100 dataset have been selected and classification has been carried out using the six copula families mentioned above. The KDD100 dataset has been trained on an HP-Z840 workstation with 10 cores, 2 x Intel CPUs (Xeon E52687Wv3), 64GB Ram and Quadro P5000 GPU. The trainings have been conducted in the MATLAB environment. In the classification phase, the 10-fold cross-validation method has been used. Since the degree of freedom (v) of the Student's-t copula is too large, an error has occurred while calculating the performance measurement. Therefore, the performance of the Student's-t copula has not shown in Table 13 and Table 14, Table 15. In Table 12 has been shown the numbers and names of the features that achieve the best performance on the KDD100 dataset.

Table 12. Numbers and names of features that achieve the best performance on the KDD100 dataset

Feature Number	Feature Name
23	count
6	dst_bytes
1	duration
32	dst_host_count
5	src_bytes
24	srv_count
33	dst_host_srv_count
4	flag

In Table 13, the performance rates of copula families on the KDD100 dataset have been shown by using the "23 6 1 32 5 24 33 4" features.

Table 13. Performance rates of copula families on KDD100 dataset

Copula Family	Method	TP	TN	FP	FN	Accuracy (%)	Features Used
independent	IFM	932928	3905436	20214	39853	98.77	"23 6 1 32 5 24 33 4"
gaussian	**IFM**	**972741**	**3896885**	**28765**	**40**	**99.41**	**"23 6 1 32 5 24 33 4"**
clayton	IFM	872190	3909171	16479	100591	97.61	"23 6 1 32 5 24 33 4"
frank	IFM	906564	3906204	19446	66217	98.25	"23 6 1 32 5 24 33 4"
gumbel	IFM	933022	3905382	20268	39759	98.77	"23 6 1 32 5 24 33 4"

As seen in Table 13, the Gaussian copula family has obtained the best success rate with 99.41% by using the "23 6 1 32 5 24 33 4" features and the IFM method. The worst success rate has been 97.61% using the "23 6 1 32 5 24 33 4" features and the IFM method for the Clayton copula family. Gaussian copula family and IFM method should be preferred for the "23 6 1 32 5 24 33 4" features in the KDD100 dataset. In Table 14, the performance rates of copula families on the KDD100 dataset have been shown by using the "23 6 1 32 5 24 33 4 3" features.

Table 14. Performance rates of copula families on KDD100 dataset using "23 6 1 32 5 24 33 4 3" features

Copula Family	Method	TP	TN	FP	FN	Accuracy (%)	Features Used
independent	IFM	933635	3905622	20028	39146	98.79	"23 6 1 32 5 24 33 4 3"
gaussian	**IFM**	**972676**	**3894142**	**31508**	**105**	**99.35**	**"23 6 1 32 5 24 33 4 3"**
clayton	IFM	831992	3912384	13266	140789	96.86	"23 6 1 32 5 24 33 4 3"
frank	IFM	906560	3906189	19461	66221	98.25	"23 6 1 32 5 24 33 4 3"
gumbel	IFM	933740	3905665	19985	39041	98.80	"23 6 1 32 5 24 33 4 3"

As seen in Table 14, the Gaussian copula family has obtained the best success rate with 99.35% by using the "23 6 1 32 5 24 33 4 3" features and the IFM method. The worst success rate has been 96.86% using the "23 6 1 32 5 24 33 4 3" features and the IFM method for the Clayton copula family. Gaussian copula family and IFM method should be preferred for the "23 6 1 32 5 24 33 4 3" features in the KDD100 dataset. In Table 15, the performance rates of copula families on the KDD100 dataset have been shown using the "23 6 1 32 5 24 33 4 3 2 7 9 10 11" features.

Table 15. Performance rates of copula families on KDD100 dataset

Copula Family	Method	TP	TN	FP	FN	Accuracy (%)	Features Used
independent	IFM	922035	3907600	18050	50746	98.60	"23 6 1 32 5 24 33 4 3 2 7 9 10 11"
gaussian	**IFM**	**972757**	**3895289**	**30361**	**24**	**99.38**	**"23 6 1 32 5 24 33 4 3 2 7 9 10 11"**
clayton	IFM	910658	3908739	16911	62123	98.39	"23 6 1 32 5 24 33 4 3 2 7 9 10 11"
frank	IFM	658964	3909941	15709	313817	93.27	"23 6 1 32 5 24 33 4 3 2 7 9 10 11"
gumbel	IFM	926372	3907381	18269	46409	98.68	"23 6 1 32 5 24 33 4 3 2 7 9 10 11"

As seen in Table 15, the Gaussian copula family has obtained the best success rate with 99.38% by using the "23 6 1 32 5 24 33 4 3 2 7 9 10 11" features and the IFM method. The worst success rate has been 93.27%, and Frank copula family has been obtained by using "23 6 1 32 5 24 33 4" features and IFM method. Gaussian copula family and IFM method should be preferred for the "23 6 1 32 5 24 33 4 3 2 7 9 10 11" features in the KDD100 dataset.

4. Conclusion and Suggestions

In this study, attack detection was performed using copula-based classifiers. In Table 16 was shown the performance comparison of copula-based classifiers for different dataset amounts

Table 16. Performance comparison of copula-based classifiers for different dataset amounts

Best Copula Algorithm		Data Set Used	Accuracy (%)
Gumbel	IFM	KDD10 (1%)	98.73
Independent Clayton Gumbel Student's-t	IFM	KDD10 (5%)	99.12
Gaussian Frank Independent Gaussian Frank Clayton Gumbel Student's-t	IFM	KDD10 (10%)	99.08
Gumbel	CML	KDD10 (50%)	98.97
Gaussian	IFM	KDD10 (100%)	98.68
Gaussian	**IFM**	**KDD100**	**99.41**

As can be seen in Table 16, attack detection was made using copula-based classifiers according to various versions of the dataset. When the amount of data set is small, the Gumbel copula-based classifier achieves better performance, while the Gaussian copula-based classifier comes to the fore as the amount of data increases. In Table 17, the success rates of some studies on IDSs in the literature and the performance of the copula-based classifiers used in this study were compared.

Table 17. Comparing the proposed study's performance to earlier IDS studies published in the literature

Some Studies in the Literature	Method Used	Data Set Used	Accuracy (%)
A.Dastanpour et al.[10]	GA+ANN	KDD'99	100.00
J.Esmaily et al. [13]	ANN	KDD'99	99.71
This study	**Copula**	**KDD'99**	**99.41**
W.Wang et al. [11]	PCA	DARPA	98.80
Y.B.Bhavsar et al. [14]	SVM	NSL-KDD	98.57
Ş.Sağıroğlu et al. [4]	ANN	KDD'99	97.92
B.W.Masduki et al. [2]	SVM	KDD'99	96.08
G.Poojitha et al. [15]	ANN	KDD'99	94.93
S.Kumar et al. [12]	ANN	KDD'99	91.90
H.A.Sonawane et al. [8]	NN	KDD'99	90.20
M.Govindarajan et al. [9]	RBF+SVM	NSL-KDD	85.19
B.Huyot et al. [5]	Copula	DARPA	79.00

As can be seen in Table 17, attack detection was carried out using many different methods in IDSs. A.Dastanpour et al. [10] achieved 100% success by using 18 features of KDD'99 data set in their study. J.Esmaily et al. [13] achieved 99.71% success by using all of 41 features in their study. In this study, as seen in Table 17, 99.41% success was achieved by using fewer (8) features then others. When the results obtained from copula-based classifiers are compared with previous studies, quite remarkable results were obtained. Thus, it was shown that copula-based classifiers can be an alternative to machine learning classifiers. As a result; In this study, Independent, Clayton, Frank, Gaussian, Gumbel and Student's-t copula-based classifiers have been preferred, and the usability of these copula-based classifiers in intrusion detection systems were investigated. Classification was performed on KDD10 (10%) and KDD100 (full) datasets of KDD'99 using copula-based classifiers. The 10-fold cross-validation method has been used in the classification phase. While all copula classifiers achieved a good 99.12% performance on the KDD10 dataset, the Gaussian copula-based classifier achieved the best success rate of 99.41% on the KDD100 dataset. As can be seen in Table 17, copula-based classifiers achieved good values when compared to other methods.

In future studies, in addition to these copula families, attack detection performances will be examined by using different copula families. In addition, the usability of ANN and copula-based approaches will be investigated.

Contributions of the authors

Mehmet Burukanlı: Methodology, Conceptualisation, Validation, Data curation, Writing original draft. Musa Çıbuk: Methodology, Conceptualisation, Validation, Investigation, Supervision.

Conflict of Interest Statement

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The study is complied with research and publication ethics.

References

[1] M. Burukanlı, Ü. Budak, and M. Çıbuk, "Saldırı Tespit Sistemlerinde Makine Öğrenme Metotlarının Kullanımı," in *Uluslararası Bilim ve Mühendislik Sempozyumu*, 2019, pp. 1052–1057.

[2] B. W. Masduki, K. Ramli, F. A. Saputra, and D. Sugiarto, "Study on Implementation of Machine Learning Methods Combination for Improving Attacks Detection Accuracy on Intrusion Detection System (IDS)," in *2015 International Conference on Quality in Research (QiR)*, 2015, pp. 56–64.

[3] M. Burukanlı, "Copula fonksiyonlarını kullanarak bilgisayar ağlarında saldırı tespiti," M.S. thesis, Bitlis Eren Üniversitesi Lisansüstü Eğitim Enstitüsü, Bitlis, Turkey, 2020.

[4] Ş. Sağıroğlu, E. N. Yolaçan, and U. Yavanoğlu, "Zeki Saldırı Tespit Sistemi Tasarımı ve Gerçekleştirilmesi," *Journal of Faculty of Engineering and Architecture of Gazi University*, vol. 26, no. 2, pp. 325–340, 2011.

[5] B. Huyot, Y. Mabiala, and J.-F. Marcotorchino, "Online Unsupervised Anomaly Detection in Large Information Systems Using Copula Theory," in *2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems*, Nov. 2014, pp. 679–684, doi: 10.1109/CCIS.2014.7175820.

[6] R. Salinas-Gutiérrez, A. Hernández-Aguirre, M. J. J. Rivera-Meraz, and E. R. Villa-Diharce, "Using Gaussian Copulas in Supervised Probabilistic Classification," in *Soft Computing for Intelligent Control and Mobile Robotics*, C. Castillo, J. Kacprzyk, and W. Pedrycz, Eds., Springer-Verlag Berlin and Heidelberg GmbH & Co. KG, 2010, pp. 355–372.

[7] M. Scavnicky, "A study of Applying Copulas in Data Mining," M.S. thesis, Charles University in Prague Faculty of Mathematics and Physics, Prague, Czech Republic, 2013.

[8] H. A. Sonawane and T. M. Pattewar, "A Comparative Performance Evaluation of Intrusion Detection Based on Neural Network and PCA," in *2015 International Conference on Communications and Signal Processing (ICCSP)*, 2015, pp. 841–845, doi: 10.1109/ICCSP.2015.7322612.

[9] M. Govindarajan and R. M. Chandrasekaran, "Intrusion Detection using an Ensemble of Classification Methods," in *Lecture Notes in Engineering and Computer Science*, 2012, vol. 1, pp. 459–464.

[10] A. Dastanpour, S. Ibrahim, R. Mashinchi, and A. Selamat, "Comparison of Genetic Algorithm Optimization on Artificial Neural Network and Support Vector Machine in Intrusion Detection System," in *2014 IEEE Conference on Open Systems (ICOS)*, Oct. 2014, pp. 72–77, doi: 10.1109/ICOS.2014.7042412.

[11] W. Wang and R. Battiti, "Identifying Intrusions in Computer Networks with Principal Component Analysis," in *First International Conference on Availability, Reliability and Security (ARES'06)*, 2006, pp. 270–279, doi: 10.1109/ARES.2006.73.

[12] S. Kumar and A. Yadav, "Increasing Performance Of Intrusion Detection System Using Neural Network," in *2014 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, 2014, pp. 546–550.

[13] J. Esmaily, R. Moradinezhad, and J. Ghasemi, "Intrusion Detection System Based on Multi-Layer Perceptron Neural Networks and Decision Tree," in *2015 7th Conference on Information and Knowledge Technology (IKT)*, May 2015, pp. 1–5, doi: 10.1109/IKT.2015.7288736.

[14] Y. B. Bhavsar and K. C. Waghmare, "Intrusion Detection System using Data Mining Technique: Support Vector Machine," *International Journal of Emerging Technologies and Advanced Engineering*, vol. 3, no. 3, pp. 581–586, 2013.

[15] G. Poojitha, K. N. Kumar, and P. J. Reddy, "Intrusion Detection using Artificial Neural Network," in *2010 Second International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Jul. 2010, pp. 1–7, doi: 10.1109/ICCCNT.2010.5592568.

[16] S. Sathe, "A Novel Bayesian Classifier using Copula Functions," *arXiv Preprint cs/0611150*, 2006.

[17] D. Qian et al., "Drowsiness Detection by Bayesian-Copula Discriminant Classifier Based on EEG Signals during Daytime Short Nap," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 4, pp. 743–754, 2017, doi: 10.1109/TBME.2016.2574812.

[18] L. Slechan and J. Górecki, "On the Accuracy of Copula-Based Bayesian Classifiers: An Experimental Comparison with Neural Networks," in *Computational Collective Intelligence*, M. Nunez, N. T. Nguyen, D. Camacho, and B. Trawinski, Eds., Springer International, Madrid, 2015, pp. 485–493.

[19] Y. Chen, "A Copula-Based Supervised Learning Classification for Continuous and Discrete Data," *Journal of Data Science*, vol. 13, pp. 769–790, 2014.

[20] N. Hammami, M. Bedda, and N. Farah, "Probabilistic Classification Based on Gaussian Copula for Speech Recognition: Application to Spoken Arabic Digits," in *Signal Processing - Algorithms, Architectures, Arrangements, and Applications Conference Proceedings (SPA)*, 2013, pp. 312–317.

[21] Y. He, J. Deng, and H. Li, "Short-Term Power Load Forecasting with Deep Belief Network and Copula Models," in *2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Aug. 2017, vol. 1, pp. 191–194, doi: 10.1109/IHMSC.2017.50.

[22] R. B. Nelsen, *An Introduction to Copulas*. Springer Science+Business Media, Inc., 2006.

[23] J. Lu, W. Tian, and P. Zhang, "The Archimedean Copulas Measure of the Risk Characteristic for the Tail Dependent Asset Returns," in *2008 International Conference on Management Science and Engineering 15th Annual Conference Proceedings*, Sep. 2008, pp. 173–181, doi: 10.1109/ICMSE.2008.4668912.

[24] P. Embrechts, F. Lindskog, and A. McNeil, "Modelling Dependence with Copulas and Applications to Risk Management," in *Handbook of Heavy Tailed Distributions in Finance*, S. T. Rachev, Ed., Elsevier, Amsterdam, 2003, pp. 329–384.

[25] T. Schmidt, "Coping with Copulas," in *Copulas: From Theory to Application in Finance*, J. Rank, Ed., Risk Books Publishing, Berkeley, 2006, pp. 3–34.

[26] E. Bouyé, V. Durrleman, A. Nikeghbali, G. Riboulet, and T. Roncalli, *Copulas for Finance-A Reading Guide and Some Applications*, SSRN Electronic Journal, 2000.

[27] A. Surana and A. Pinto, "Analysis of Stochastic Automata Networks Using Copula Functions," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sep. 2010, pp. 1699–1706, doi: 10.1109/ALLERTON.2010.5707121.

[28] B. Z. Karagül, "Hayat Dışı Sigortalarda Doğrusal Olmayan Bağımlılığın Kopulalar ile Dinamik Finansal Analizi," M.S. thesis, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara, Turkey, 2013.

[29] G. Yapakçı, "Kopulalar Teorisinin Finansta Uygulaması," M.S. thesis, Ege Üniversitesi Fen Bilimleri Enstitüsü, İzmir, Turkey, 2007.

[30] S. Aslan, S. Çelebioğlu, and F. Öztürk, "İki Boyutlu Arşimedyen Kopulalarda İstatistiksel Sonuç Çıkarımı ve Bir Uygulama," *Gazi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, vol. 14, no. 2, pp. 1–18, 2012.

[31] L. Andersen and J. Sidenius, "Extensions to the Gaussian Copula: Random Recovery and Random Factor Loadings," *Journal of Credit Risk*, vol. 1, no. 1, pp. 29–70, 2005, doi: 10.21314/jcr.2005.003.

[32] D. Çatal and R. S. Albayrak, "Riske Maruz Değer Hesabında Karışım Kopula Kullanımı: Dolar-Euro Portföyü," *Yaşar Üniversitesi E-Dergi*, vol. 8, no. 31, pp. 5187–5202, 2013.

[33] M. Mehdizadeh, R. Ghazi, and M. Ghayeni, "Power System Security Assessment with High Wind Penetration Using the Farms Models Based on Their Correlation," *IET Renewable Power Generation*, vol. 12, no. 8, pp. 893–900, 2018, doi: 10.1049/iet-rpg.2017.0386.

[34] P. Hájek and R. Mesiar, "On Copulas, Quasicopulas and Fuzzy Logic," *Soft Computing*, vol. 12, no. 12, pp. 1239–1243, 2008, doi: 10.1007/s00500-008-0286-z.

[35] J. Yan, "Enjoy the Joy of Copulas: With a Package copula," *Journal of Statistical Software*, vol. 21, no. 4, pp. 1–21, 2007, doi: 10.18637/jss.v021.i04.

[36] H. He and P. K. Varshney, "A Coalitional Game for Distributed Inference in Sensor Networks with Dependent Observations," *IEEE Transactions on Signal Processing*, vol. 64, no. 7, pp. 1854–1866, 2016, doi: 10.1109/TSP.2015.2508781.

[37] G. Van Der Wulp, "Using Copulas in Risk Management," M.S. thesis, Tilburg University Department of Econometrics, Tilburg, Netherlands, 2003.

[38] M. Sklar, "Fonctions de Répartition à n Dimensions et Leurs Marges," *Publications de l'Institut Statistique de l'Université de Paris*, vol. 8, pp. 229–231, 1959.

[39] A. Alhan, "Bağımsızlık Kapulasını İçeren Kapula Aileleri, Kapula Tahmin Yöntemleri ve İstanbul Menkul Kıymetler Borsasında Sektörler Arası Bağımlılık Yapısı," Ph.D. dissertation, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara, Turkey, 2008.

[40] A. M. Karataş, "Modeling of Daily Maximum and Minimum Temperature Changes in Bitlis Province Using Copula Method," *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, vol. 7, no. 2, pp. 268–275, 2018.

[41] J. Lu, W.-J. Tian, and P. Zhang, "The Extreme Value Copulas Analysis of the Risk Dependence for the Foreign Exchange Data," in *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing (WiCom)*, Oct. 2008, pp. 1–6, doi: 10.1109/WiCom.2008.2405.

[42] P. Mou, F. Tao, C. Jia, and W. Ma, "A Copula-Based Function Model in Fuzzy Reliability Analysis on the Planetary Steering Gear," in *2013 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE)*, Jul. 2013, pp. 375–378, doi: 10.1109/QR2MSE.2013.6625605.

[43] A. M. Karakaş, "Modelling Temperature Measurement Data by Using Copula Functions," *Bitlis Eren Üniversitesi Fen Bilimleri ve Teknoloji Dergisi*, vol. 7, no. 1, pp. 27–32, 2017.

[44] S. Jadhav and R. Daruwala, "3-D Modeling of Statistical Dependencies Using Copulas for Wireless Sensor Network," in *Proceedings of the 2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2016, pp. 1886–1889, doi: 10.1109/WiSPNET.2016.7566469.

[45] C. D. Tran, O. O. Rudovic, and V. Pavlovic, "Unsupervised Domain Adaptation with Copula Models," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2017, pp. 1–6.

[46] C. Romano, "Calibrating and Simulating Copula Functions: An Application to the Italian Stock Market," *Risk Management Functional Capital*, vol. 180, pp. 1–26, 2002.

[47] Anonim, "Probability Distributions," 2017. [Online]. Available: http://www.nematrian.com/Pages/ProbabilityDistributionsCombined.pdf. [Accessed: Apr. 13, 2020].

[48] E. E. Sezgin, "Finansal Bağımlılık Analizi: Vine ve CD Vine Copula Yaklaşımları," M.S. thesis, Bitlis Eren Üniversitesi ve Fırat Üniversitesi Fen Bilimleri Enstitüsü, Bitlis, Turkey, 2019.

[49] S. Arslan, "Arşimedyen Kapulalar Üzerine Bir Çalışma," Ph.D. dissertation, Ankara Üniversitesi Fen Bilimleri Enstitüsü, Ankara, Turkey, 2013.

[50] S. Çelebioğlu, "Arşimedyen Kapulalar ve Bir Uygulama," *Selçuk Üniversitesi Fen Fakültesi Fen Dergisi*, vol. 22, no. 1, pp. 43–52, 2003.

[51] S. S. Galiani, "Copula Functions and Their Application in Pricing and Risk Managing Multiname Credit Derivative Products," M.S. thesis, Department of Mathematics, King's College London, London, UK, 2003.

[52] H. Manner, *Estimation and Model Selection of Copulas with an Application to Exchange Rates*. Maastricht: Maastricht Research School of Economics of Technology and Organizations (METEOR) Press, 2007.

[53] Y. Dong, S. Zhang, G. Fan, L. Zhang, L. Yi, and M. Lin, "Application of Copula Function in the Reliability Analysis of the Electrical System and the Power Device of Certain-Type Armored Vehicle," in *CSAE 2012 - Proceedings of the 2012 IEEE International Conference on Computer Science and Automation Engineering*, 2012, vol. 1, pp. 386–389, doi: 10.1109/CSAE.2012.6272621.

[54] A. Setiawan, Soeheri, E. Panggabean, M. A. Elhias, F. Ikorasaki, and B. Riski, "Efficiency of Bayes Theorem in Detecting Early Symptoms of Avian Diseases," in *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, Aug. 2018, pp. 1–5, doi: 10.1109/CITSM.2018.8674273.

[55] N. S. B. Sembiring, E. Ginting, M. Fauzi, Yudi, F. Tambunan, and E. V. Haryanto, "An Expert System to Diagnose Herpes Zoster Disease Using Bayes Theorem," in *2019 7th International Conference on Cyber and IT Service Management (CITSM)*, Nov. 2019, pp. 1–3, doi: 10.1109/CITSM47753.2019.8965381.

[56] A. H. Jahromi and M. Taheri, "A Non-Parametric Mixture of Gaussian Naive Bayes Classifiers Based on Local Independent Features," in *2017 Artificial Intelligence and Signal Processing Conference (AISP)*, Oct. 2017, vol. 2018, pp. 209–212, doi: 10.1109/AISP.2017.8324083.

[57] K. P. Murphy, "Naive Bayes Classifiers," 2006. [Online]. Available: https://www.ic.unicamp.br/~rocha/teaching/2011s1/mc906/aulas/naive-bayes.pdf. [Accessed: Apr. 02, 2020].

[58] K. Netti and Y. Radhika, "A Novel Method for Minimizing Loss of Accuracy in Naive Bayes Classifier," in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Dec. 2015, pp. 1–4, doi: 10.1109/ICCIC.2015.7435801.

[59] F.-J. Yang, "An Implementation of Naive Bayes Classifier," in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, Dec. 2018, pp. 301–306, doi: 10.1109/CSCI46756.2018.00065.

[60] Anonim, "Maximum A Posteriori Estimation," 2020. [Online]. Available: https://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation. [Accessed: May 04, 2020].

[61] H. Bogunovic, J. M. Pozo, R. Cardenes, L. S. Roman, and A. F. Frangi, "Anatomical Labeling of the Circle of Willis Using Maximum A Posteriori Probability Estimation," *IEEE Transactions on Medical Imaging*, vol. 32, no. 9, pp. 1587–1599, 2013.

[62] F. Peng, D. Schuurmans, and S. Wang, "Augmenting Naive Bayes Classifiers with Statistical Language Models," *Information Retrieval Boston*, vol. 7, no. 3–4, pp. 317–345, 2004.

[63] M. J. Islam, Q. M. J. Wu, M. Ahmadi, and M. A. Sid Ahmed, "Investigating the Performance of Naive Bayes Classifiers and K-Nearest Neighbor Classifiers," in *2007 International Conference on Convergence Information Technology (ICCIT)*, Nov. 2007, pp. 1541–1546, doi: 10.1109/ICCIT.2007.148.

[64] J.-H. Lee, J.-H. Lee, S.-G. Sohn, J.-H. Ryu, and T.-M. Chung, "Effective Value of Decision Tree with KDD 99 Intrusion Detection Datasets for Intrusion Detection System," in *2008 10th International Conference on Advanced Communication Technology (ICACT)*, Feb. 2008, vol. 2, pp. 1170–1175, doi: 10.1109/ICACT.2008.4493974.

[65] Anonim, "Kddcup1999," 1999. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/task.html. [Accessed: Apr. 09, 2020].

[66] B. Nethu, "Classification of Intrusion Detection Dataset Using Machine Learning Approaches," *International Journal of Electronics and Computer Science Engineering*, vol. 1, no. 3, pp. 1044–1051, 2012.

[67] D. H. Deshmukh, T. Ghorpade, and P. Padiya, "Intrusion Detection System by Improved Preprocessing Methods and Naive Bayes Classifier Using NSL-KDD 99 Dataset," in *2014 International Conference on Electronics and Communication Systems (ICECS)*, Feb. 2014, pp. 1–7, doi: 10.1109/ECS.2014.6892542.

[68] T. Janarthanan and S. Zargari, "Feature Selection in UNSW-NB15 and KDDCUP'99 Datasets," in *2017 IEEE 26th International Symposium on Industrial Electronics (ISIE)*, 2017, pp. 1881–1886.

[69] G. Meena and R. R. Choudhary, "A Review Paper on IDS Classification Using KDD 99 and NSL KDD Dataset in WEKA," in *2017 International Conference on Computer, Communications and Electronics (Comptelix)*, Jul. 2017, pp. 553–558, doi: 10.1109/COMPTELIX.2017.8004032.

[70] Y. Vural, "Kurumsal Bilgi Güvenliğinde Güvenlik Testleri ve Öneriler," *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, vol. 26, no. 1, pp. 89–103, 2011.

[71] Anonim, "The UCI KDD Archive Information and Computer Science University of California, Irvine," 1999. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. [Accessed: Jan. 26, 2021].

[72] M. Burukanlı, M. Çıbuk, and Ü. Budak, "Saldırı Tespiti için Makine Öğrenme Yöntemlerinin Karşılaştırmalı Analizi," *BEÜ Fen Bilimleri Dergisi*, vol. 10, no. 2, pp. 613–624, 2021.

[73] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005, doi: 10.1109/TPAMI.2005.159.

[74] M. Çıbuk, U. Budak, Y. Guo, M. C. Ince, and A. Sengur, "Efficient Deep Features Selections and Classification for Flower Species Recognition," *Measurement*, vol. 137, pp. 7–13, Apr. 2019, doi: 10.1016/j.measurement.2019.01.041.

[75] S. Wang, Y.-H. Zhang, J. Lu, W. Cui, J. Hu, and Y.-D. Cai, "Analysis and Identification of Aptamer-Compound Interactions with a Maximum Relevance Minimum Redundancy and Nearest Neighbor Algorithm," *BioMed Research International*, vol. 2016, pp. 1–12, 2016. [Online]. Available: http://downloads.hindawi.com/journals/bmri/2016/8351204.pdf. [Accessed: Apr. 29, 2020].