TÜRKİYE CUMHURİYETİ'NİN YÜZÜNCÜ YILI

PRACTICAL ARTIFICAL INTELLIGENCE APPLICATIONS - I

Editör

Eyyüp Gülbandılar

BİDGE

BIDGE Publications

Practical Artificial Intelligence Applications 1

Editor: Prof.Dr.Eyyüp GÜLBANDILAR

ISBN: 978-625-372-393-4

Page Layout: Gözde YÜCEL

1st Edition:

Publication Date: 25.12.2024

BIDGE Publications,

Certificate No: 71374

www.bidgeyayinlari.com.tr - bidgeyayinlari@gmail.com

Krc Bilişim Ticaret ve Organizasyon Ltd. Şti.

Güzeltepe Mahallesi Abidin Daver Sokak Sefer Apartmanı No: 7/9 Çankaya / Ankara

BİDGE Yayınları

Content

CHAPTER I

Impact Of Data Balancing Methods On Classifier Performance: Classification Of Metaverse-Related X (Twitter) Data

Gül Cihan HABEK[1]

Introduction

X, formerly known as Twitter, is a social media platform that serves as a significant source of data for popular topics and trends. Millions of X users share real-time thoughts and feelings, as well as text, photos, and videos related to current events, trends, societal issues, or commercial products in what are known as "tweets." These shared tweets create large datasets on any given topic (Habek, 2022). Users can indicate the topics of their shared tweets using hashtags, enabling the data to be quickly categorized. The shared tweets are frequently used in marketing strategy, campaign analysis, or sentiment analysis studies.

[1] Res. Asc., Karamanoglu Mehmetbey University, Faculty of Engineering, Department of Computer Engineering, Karaman/Turkey, Orcid: 0000-0003-1748-3486, gulhabekkmu@gmail.com

The Metaverse is a concept where people can interact in a virtual world independently of their physical location, using the internet and physical equipment (Wang et al., 2022). This concept, representing a virtual universe, is composed of a combination of various technologies such as social media, Augmented Reality (AR), Virtual Reality (VR), cryptocurrencies, and online gaming (Cheng, 2023). In this study, sentiment analysis is aimed to be conducted on tweets related to the Metaverse shared on the X platform, where users can come together, interact, play games, explore virtual worlds, and acquire virtual properties within the Metaverse world.

User-generated content related to the Metaverse tends to carry a predominantly positive sentiment, with negative tweets being in the minority. In studies like this, datasets where one class has significantly more instances than other classes are referred to as imbalanced datasets. Machine learning models built on imbalanced datasets often struggle to adequately learn from the minority class data while better learning from the majority class, leading to challenges in predicting data from the minority class (He & Garcia, 2009). In many cases, the high number of correct predictions in the majority class can lead to a seemingly high model accuracy, but this success may be far from true accuracy. This situation often hinders reaching solutions in real-world problems where a balanced distribution is lacking. For example, in tasks such as detecting rare diseases in medical diagnoses, fraud detection, customer churn analysis, identifying rare attacks in cybersecurity, or detecting faulty products in quality control, accurate prediction of the minority class is crucial. Even though the results obtained in these studies may appear successful, they can be far from reality.

Many classification models, unlike real-world data, operate under the assumption that the dataset is balanced, and there are various data balancing methods in the literature to address this issue. These methods are divided into three classes: oversampling to increase the number of examples in the minority class, under sampling to reduce the number of examples in the majority class, and applying hybrid methods to achieve a more balanced data distribution.

In a study conducted by Stando et al. (Stando, Cavus, & Biecek, 2024) in the literature, they aimed to investigate the effects of data balancing methods on model performance. In the study, simulated and real datasets were balanced using various data-balancing methods, and their performances were compared. The study concluded that the greatest performance improvement was achieved when using the Random Undersampling (RUS) method.

In another study, Dina et al. (Dina, Siddique, & Manivannan, 2022) aimed to classify two different intrusion detection datasets, NSL-KDD and UNSW-NB15, using Decision Tree (DT), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF), Feedforward Neural Network (FNN), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) classifiers. Both datasets used in the study were imbalanced, so the data was balanced using Random Over Sampling (ROS) and Conditional Generative Adversarial Network (CTGAN) data balancing methods, and the performance changes on the classifiers were examined. According to the results of the study, the prediction accuracy of various machine learning classifiers trained on the balanced datasets with synthetic samples generated by CTGAN increased by up to 8% compared to the same classifiers trained on imbalanced data. However, when trained on data balanced with ROS, the accuracy decreased.

Ebrahimy et al. (Ebrahimy, Mirbagheri, Matkan, & Azadbakht, 2022) conducted a study addressing the issue of class imbalance that occurs when the number of correctly classified pixels is greater than the number of incorrectly classified pixels. In their study, they compared the performance of three ensemble machine learning algorithms—Random Forest (RF), Rotation Forest (RoF), and Stochastic Gradient Boosting (SGB)—using three different data balancing methods on a remote sensing image dataset: Random Over Sampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN). The experimental results showed that the combination of the SMOTE data balancing method and the RF classifier significantly improved the performance.

In their study, Jadhav et al. (Jadhav, Samih, Elmannai, & Karim, 2022) aimed to evaluate the performance of six different data balancing methods—Under Sampling, Over Sampling, Hybrid Sampling, Random Over Sampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), and Clustering-Based Under Sampling (CBUS). The study utilized twenty-five different datasets. The results obtained using Decision Tree (C4.5), K-Nearest Neighbors (KNN), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM) classifiers indicated that data balancing methods improved the classification performance.

Tolba et al. (Tolba, Ouadfel, & Meshoul, 2021) conducted a study to automatically detect harassing tweets sent from X. The imbalanced dataset used in the study was balanced using various methods including SMOTE, NearMiss (NM), Cost-Sensitive Learning, KMeans SMOTE, TextGan, LoRAS, SDS, and Cluster-Based Under Sampling, and the performances were compared.

Duan et al. (Duan, Wei, Liu, & Yin, 2020) proposed a new ensemble classifier model based on KMeans and a resampling technique (EKR) to reduce the imbalance ratio in datasets and improve performance. The proposed model was applied to 16 imbalanced datasets, yielding effective and feasible results.

In their study, Halimu et al. (Halimu & Kasem, 2021) proposed a new data balancing method called sBal to address the imbalance issue in binary imbalanced datasets. When applied to 50 different datasets and compared with SMOTEBagging, SMOTEBoost, RUSBoost, and RAMOBoost methods, the proposed sBal method was observed to significantly improve classification performance.

In their study, Potharaju et al. (Potharaju, Sreedevi, Ande, & Tirandasu, 2019) attempted to classify the health status of fetuses in the womb using six different classification algorithms: Jrip, Ridor, J48, NBStar, Instance-Based k-Nearest Neighbors (IBk), and Kstar. The study utilized the Cardiotocography dataset, which contains measurements classified by expert obstetricians. The class imbalance in the dataset was addressed using the SMOTE data balancing method. When comparing the experimental results obtained from the imbalanced and balanced datasets, it was concluded that better performance values were achieved with the balanced dataset.

Domingues et al. (Domingues, Amorim, Abreu, Duarte, & Santos, 2018) addressed the issue of sequential data imbalance by using four different non-sequential data balancing techniques: ROS, SMOTE, CBO, and DEAGO. They proposed four different methods: Two become three (2to3), Feature by feature, Centroid based, and PCA based. These methods were applied to two different datasets for experimental comparison.

In the most recent literature study, Zhongbin et al. (Sun et al., 2015) addressed the problem of binary class imbalance. They proposed a method that transforms imbalanced binary class datasets into balanced multi-class datasets. The performance of the proposed method was evaluated on 46 highly imbalanced datasets using six different classifiers: NB, C4.5, RIPPER, RF, SMO, and IBk. When compared with traditional methods such as RUS, ROS, SMOTE, MetaCost, Bagging, Boosting, EasyEnsemble, SMOTEBoost, RUSBoost, and UnderBagging, the proposed method demonstrated superior performance.

Various studies have been conducted in the literature using data balancing methods. However, it has been observed that there are relatively few studies evaluating the performance of these methods. In this study, the impact of different data balancing methods on the performance of traditional machine learning algorithms was compared using a specially created imbalanced dataset. During the sentiment analysis phase, the performance of traditional machine learning algorithms NB, LR, SVM, DT, and RF trained with 12 balanced datasets was measured using accuracy, precision, recall, f1-score, g-mean, and ROC metrics. The experimental results showed that data balancing methods improve the performance of classification models, with the best results obtained using NB classifier with SMOTE, Borderline-SMOTE, KMeans-SMOTE oversampling methods, and the SMOTE+Tomek

Links hybrid method. Subsequently, results obtained from accuracy, precision, recall, and f1-score metrics were subjected to Friedman and Nemenyi post-hoc tests for statistical analysis, examining the relationship between the methods. The statistical analyses indicate a relationship between the performances of the used data balancing methods. Additionally, two different imbalanced datasets shared on the Kaggle platform, the binary “Amazon Product Reviews” and the eight-class “Emotion Dataset,” were balanced using the same methods, and results were obtained for six performance metrics using the same traditional machine learning algorithms.

In the subsequent sections of the study, under the "Materials and Methods" heading, detailed information was provided about the datasets used, the data balancing methods employed, the traditional machine learning algorithms utilized, and the performance metrics evaluated. The study's findings and comparative analysis were presented in the "Results and Discussion" section, while conclusions drawn from the results and suggestions for future improvements were discussed in the "Conclusion and Suggestions" section.

Material and Method

This section briefly explains the imbalanced datasets used in the study, the data balancing methods employed to address the imbalance between classes, traditional machine learning algorithms, and evaluation metrics.

1.1. Metaverse dataset

During the data collection and labeling phase, data is typically collected for a specific purpose and labeled manually or using automated methods. In this study, a sentiment analysis was conducted on the topic of the Metaverse, and all tweets extracted from platform X were manually labeled. To create the dataset, the snscrape library, commonly used in data scraping projects, was employed to extract Turkish tweets posted with the #metaverse hashtag from platform X between 10.06.2022 and 10.01.2023, resulting in a dataset of 5000 tweets (Habek, 2023). To enhance the performance and generalization capability of the sentiment analysis models, the dataset underwent the following preprocessing steps before being fed into the models.

- Words starting with # and @ symbols were removed from the data.
- All letters were converted to lowercase.
- Multiple spaces were cleaned up, and punctuation marks were removed.
- Duplicate data entries were deleted to avoid negatively impacting learning.
- Spelling errors were corrected, and words in foreign languages were translated into Turkish.
- Commonly used Turkish stopwords like “acaba”, “belki”, “lakin”, and the 2000 least frequent words were removed to reduce noise in the dataset.

After the cleaning step, 2365 unique tweets were obtained. These tweets were classified into five different categories based on their content: positive, negative, neutral, advertisement, and to be deleted, as shown in Table 1. Table 1 also includes examples of tweets from different categories.

Table 1: Tweet counts and examples in each class

Classes	Tweet counts	Tweet examples
positive	1321	"pek kişinin gözünde sanal gerçeklik vr teknolojileri birbirleriyle bağlantılı olarak ilerliyor popüler metaverse şirketleri yatırım yapıyor"
negative	662	"uyudum uyandım hala aynı yaklaşık iki haftadır aynı yerinde sayıyor bana göre ayı sezonlarının kötü tarafı düşüş değil tür bıktırma hamleleridir"
neutral	64	"kriptonun bir zamanı var sepet yaparken kategorilendirip almanız sizin lehinizedir metaverse koinler hisselerine dayalıdır yükselişi webiotpp alanda yatırım alabilir defi projeler kripto değerleme stratejilerine göre yatırım alır"
advertisement	382	"kripto ticaretinizi hızlı güvenli başarılı hale getirmek yapay zeka otomatik ticaret botuyla tanışın"
deleted	479	"ülkenin gündemine bakayım dedim cidden yazık insanlık yapay rahim gerçekleştirdi üstün zekalı kalıtsal hastalıklardan arınmış bebeklerden bahsediyorlar metaverse tasarlıyorlar hayatı eğitimi sağlığı oraya taşıyıp bunları konuşuyorlar bizim ülkenin gündemine bak yazık"

The criteria used for classifying the tweets into different categories are as follows:

- **positive**: Includes positive thoughts and expectations about Metaverse, comments on investments made by individuals and companies, and future expectations.
- **negative**: Includes negative statements about Metaverse, comments from those who do not believe in Metaverse technology, and remarks stating that it is not supported from a religious perspective.
- **neutral**: Includes users' current basket shares, news content shares, and terminological explanations.
- **advertisement**: Includes promotional content for making money through Metaverse.

- **deleted**: Includes tweets on different topics, which are planned to be removed from the dataset.

After the labeling process was completed, the “to be deleted” class, as well as the neutral and advertisement classes (due to containing very little data), were removed from the dataset as previously planned. The study continued with the remaining positive and negative classes.

1.2. Amazon product reviews dataset

The Amazon Product Reviews Dataset consists of 19,966 customer reviews written in English, divided into two classes: positive and negative sentiment. During the text preprocessing stage, all letters were converted to lowercase, and common English stopwords such as “a”, “an”, “the”, “in” were removed, along with the 2000 least frequent words from the dataset, completing the text preprocessing step.

The classes have an imbalanced distribution, and examples of data within each class after the text preprocessing step are provided in Table 2.

Table 2: Data counts and examples in each class

Classes	Tweet counts	Tweet examples
positive	15230	“best apps acording bunch people agree bombs eggs pigs tnt king pigs realustic stuff”
negative	4766	“angry people book moron would waste time playing azimuth slingshot call cultural phenomenon word oh yes moron 5 minutes played thing still scar”

2.3. Emotion dataset

Another dataset used is the Emotion dataset obtained from the Kaggle platform. This dataset consists of eight different emotion classes in English: anger, disgust, fear, joy, neutral, sadness, shame, and surprise. The reason for selecting this dataset is to compare the performance of data balancing methods on multi-class datasets in addition to binary datasets. The same text preprocessing steps applied to the Amazon Product Reviews dataset were applied to obtain clean data from this dataset as well.

Table 3 includes the number of data samples in each class of the final dataset under consideration, along with examples of data after the text preprocessing step.

Table 3: Data counts and examples in each class

Classes	Tweet counts	Tweet examples
anger	4297	"someone tried rape one best friends claimed fault hassled claimed loose threatened sue"
disgust	856	"understand mood disgruntled entitled call egm"
fear	5410	"america new north korea china germany russia military controlled selfspying censoring paranoid antagonists"
joy	11045	"seein folks overcome limitations seeing dream capacity compass"
neutral	2254	"dont irritable dont understand program thats"
sadness	6722	"brothers friends house nothing happens come back home"
shame	146	"jasper reddened pressed lip together farraline looked embarrassed"
surprise	4062	"think friends family going big near future"

2. Data balancing methods

Minority class refers to categories of data that are found in very low proportions 14itera rarely observed within a dataset (Bagui, Mink, Bagui, Subramaniam, & Wallace, 2023). Examples in the minority class are significantly fewer compared to other classes, a situation referred to as data imbalance. In the literatüre, various data balancing methods are used to address the problem of data imbalance. These methods are classified into three categories: oversampling, undersampling, and hybrid methods that combine oversampling with undersampling (Kotsiantis, Kanellopoulos, & Pintelas, 2006).

Oversampling involves generating synthetic examples from the minority class in a dataset to achieve a balanced distribution among classes. By oversampling, the representation of the minority class is increased, thus allowing the classifier model to be trained in a more balanced manner (Kubat & Matwin, 1997). Five different oversampling techniques, namely ROS, SMOTE, Borderline-SMOTE, Kmeans SMOTE, and ADASYN were used to balance the dataset.

In the ROS (Random Over-Sampling) method, the balance between classes is achieved by including copies of minority class samples or synthetic examples with similar features in the minority class. However, this method can lead to overfitting risks, where the model may perform exceptionally well on the training data but fail to generalize to new, unseen data.

SMOTE (Synthetic Minority Oversampling Technique) is a method aimed at balancing classes by generating synthetic examples from the minority class (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). In cases where the minority class is underrepresented, to reduce overfitting and improve generalization ability, SMOTE identifies the k-nearest neighbors for each example in the minority class. A difference vector is then calculated between the example and a randomly selected neighbor from these neighbors. A synthetic example is created by adding a weight value between 0 and 1 to this difference vector.

Borderline-SMOTE is a method developed to improve upon the weaknesses of the SMOTE method (Han, Wang, & Mao, 2005). The main objective of Borderline-SMOTE is to focus on minority class examples located at the borders (borderline) to help the model better learn the decision boundaries of the minority class for more accurate classification. In this method, for each minority class example, the KNN (K-Nearest Neighbors) algorithm is used to identify its nearest neighbors. A random neighbor of the minority class example is selected, and a synthetic example is generated by adding a weight to the difference between these two examples. This process continues until a specified threshold value is reached. The threshold value is based on the probability of an example belonging to the majority class. If this probability is lower than a certain threshold value, the example is augmented with synthetic examples.

Kmeans SMOTE is another type of the SMOTE algorithm. It consists of three stages: clustering, filtering, and oversampling (Douzas, Bacao, & Last, 2018). In the clustering stage, the dataset is divided into k clusters using the Kmeans clustering algorithm to group together data points with similar features. In the filtering stage, clusters with a high density of minority class examples are selected to ensure that synthetic examples represent the minority class. In the oversampling stage, SMOTE is applied to increase the number of minority class examples within the selected clusters. Unlike regular SMOTE, which randomly generates new examples, Kmeans SMOTE places synthetic examples close to cluster centers, creating a more balanced and realistic dataset. This approach helps reduce class imbalance while mitigating the risk of overfitting caused by excessive oversampling.

ADASYN (Adaptive Synthetic) is an enhanced version of the SMOTE method for data balancing. In ADASYN, while reducing class imbalance, the classification decision boundary is adjusted around difficult examples to improve learning performance (He, Bai, Garcia, & Li, 2008). In other words, in the ADASYN method, synthetic examples better mimic the characteristics of the minority class, making synthetic data more consistent with real data and enhancing the distinction between classes.

Undersampling methods aim to rebalance the dataset by reducing or removing data from the majority class to equalize class distributions (Batista, Prati, & Monard, 2004; Tahir, Kittler, Mikolajczyk, & Yan, 2009). Deleting data from the majority class can lead to the loss of important information in the dataset and a reduction in dataset size. When working with smaller datasets, models may suffer from reduced generalization ability, producing results that are less representative of real-world data. In the study, in addition to oversampling methods, four types of undersampling methods RUS (Random Under-Sampling), Tomek Links, Edited Nearest Neighbors (ENN), and NM were used to balance the dataset, aiming to compare their performances with different methods.

RUS (Random Under Sampling) aims to balance the dataset by randomly removing examples from the majority class (Saripuddin, Suliman, S.S., & Jorgensen, 2021). In the RUS (Random Under Sampling) method, a specified proportion of randomly selected examples from the majority class are removed, while all examples from the minority class are preserved.

In the CNN (Condensed Nearest Neighbor Rule) method, an example from the minority class is compared with a randomly selected example from the majority class. Examples with distances below a specified threshold value are retained, while examples with distances above the threshold are removed. This process aims to condense the dataset by preserving examples that are closer in distance across classes (Hart, 1968). This process continues until all examples are classified and no changes are made. In CNN undersampling, since initial data selection is random, there’s a higher probability that examples come from the larger majority class. This could lead to unnecessary examples being stored and inadequate representation of the minority class. Therefore, to address this concern, two modifications were made to the CNN method, leading to the development of the Tomek Links method (Tomek, 1976). The method used to determine boundaries between classes states that if two examples from different classes are each other’s nearest neighbors, they form a Tomek link. Tomek links help clarify boundaries between classes by removing such examples that connect different classes.

The Edited Nearest Neighbors (ENN) algorithm is a method that attempts to balance the dataset with minimal information loss (Wilson, 1972). In the Edited Nearest Neighbors (ENN) algorithm, each example in the minority class is reorganized by examining the classes among its neighbors. The method progresses by removing examples that are neighbors to the majority class majority, thereby reducing the influence of majority class examples on the minority class (Brownlee, 2020).

- In the NearMiss (NM) method, data points are removed based on their proximity to the minority class compared to the majority class (Mqadi, Naicker, & Adeliyi, 2021). The NearMiss method has three versions:
- NearMiss-1 aims to select examples from the minority class that are closest to the k nearest neighbors within the majority class.
- NearMiss-2 selects examples from the majority class that have the shortest average distance to the farthest examples in the minority class.
- NearMiss-3 first identifies and preserves the nearest neighbors of examples in the minority class. Then, it selects examples from the majority class based on the average distance to their nearest neighbors.

The aim was to mitigate the adverse effects of imbalanced datasets on classifier models by applying the described oversampling methods, undersampling methods, and hybrid methods (SMOTE + Tomek Links, SMOTE + ENN, and SMOTE + RUS) to three specific datasets: a custom-created imbalanced Metaverse dataset and two imbalanced datasets downloaded from Kaggle. From each of these three datasets, balanced datasets were created with 12 versions each. The performance of these methods was then compared using various classifiers.

3. Machine learning algorithms

In the classification stage of the study, traditional machine learning algorithms including NB (Naive Bayes), LR (Logistic Regression), SVM (Support Vector Machine), DT (Decision Tree), and RF (Random Forest) algorithms were utilized.

Naïve Bayes (NB) algorithm takes its name from Thomas Bayes. This method, which is simple and easy to understand, is commonly preferred for classification problems. It is a supervised learning method based on probabilistic calculations to determine the class of examples in the dataset (Dewi, Chen, Christanto, & Cauteruccio, 2023).

Logistic Regression (LR) is a classification algorithm, despite its name. It can be used for binary or multi-class problems. This algorithm is a probabilistic model based on the logistic function.

Support Vector Machine (SVM) is a supervised learning algorithm that performs well on small and medium-sized datasets (Nurkholis, Alita, & Munandar, 2022). It uses a maximum-margin line or hyperplane to separate two classes. The boundary of the line is called the "margin," and a wider margin leads to higher classification accuracy.

Decision Tree (DT) is a supervised learning algorithm that works with a root-node-tree structure. It is used for both classification and regression problems.

Random Forest (RF) is a supervised learning algorithm that obtains results using multiple decision trees. The decision made by the majority is considered the final outcome. It is used for classification and regression problems.

4. Evaluation Metrics

The performance of machine learning algorithms is evaluated using a confusion matrix. In two-class classification problems, the confusion matrix is presented in Table 4. Columns in Table 4 represent the actual class labels, while rows represent the predicted class labels.

Table 4: Confusion matrix

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

TP (True Positives) and TN (True Negatives) in the error matrix indicate the number of correctly classified positive and negative examples; FP (False Positives) and FN (False Negatives) give the number of misclassified positive and negative samples.

In the study, three different data sets balanced with various data balancing methods were classified with traditional machine learning algorithms and their performances were evaluated using accuracy, precision, recall, F1-score, ROC AUC, and G-mean metrics. Brief descriptions and formulas of the metrics used are given below.

- **Accuracy:** The metric that gives the rate of correct classification of examples. It is calculated by the ratio of the total number of correct predictions of the model to the total number of examples.

$$acc\colon\ (TP+TN)/Total \tag{1}$$

- **Precision:** This metric gives the ratio of how many of the examples predicted as positive were correctly predicted as positive.

$$precision\colon TP/(TP+FP) \tag{2}$$

- **Recall:** This metric gives the rate of correctly predicting true positives.

$$recall\colon TP/(TP+FN) \tag{3}$$

- **F1-score:** The harmonic means of sensitivity and recall metrics, which are considered rival metrics in literature, gives the F1-score value.

$$f1-score\colon\ (2*precision*recall)/(precision+recall) \tag{4}$$

- **ROC curve and AUC:** ROC (Receiver Operating Characteristic) curve is a frequently used metric to measure the performance of a classification model. Equation 5 illustrates the relationship between TPR (True Positive Rate) and FPR (False Positive Rate) (Muschelli, 2020). The area under the curve (AUC) is a measure of how well models can distinguish between classes. As performance improves, the ROC curve approaches the point (0,1) and AUC approaches 1.

$$TPR = TP/(TP + FN);\ \ FPR = FP/(FP + TN) \qquad (5)$$

$$AUC = \sum_{i=1}^{n}(FPR_{i+1} - FPR_i) * (TPR_{i+1} + TPR_i)/2 \qquad (6)$$

- **G-mean:** G-mean (Geometric Mean) is a metric particularly used to evaluate the performance of imbalanced datasets. As seen in Equation 7, it gives the geometric mean of the classification successes of the minority and majority classes (Góra & Skowron, 2023).

$$G - mean = \sqrt{((TP/(TP + FN)) + (TN/(TN + FP)))} \qquad (7)$$

Results and Discussion

The experimental results from the study are categorized under two main headings: Data Balancing Results and Classification Results.

1. Data Balancing Results

The study addressed three separate datasets with different distributions and numbers of instances, attempting to balance them using various data balancing methods. Effective methods for each dataset were analyzed. The study specifically calculated the coefficient of variance by taking the ratio of the standard deviation of class sample numbers to their mean and calculated the imbalance ratio by taking the ratio of the maximum number of samples in a class to the minimum number. Class distributions, coefficient of variance, and imbalance ratios are shown in Figures 1, 2, and 3 respectively, both in their imbalanced states and after applying various balancing methods.

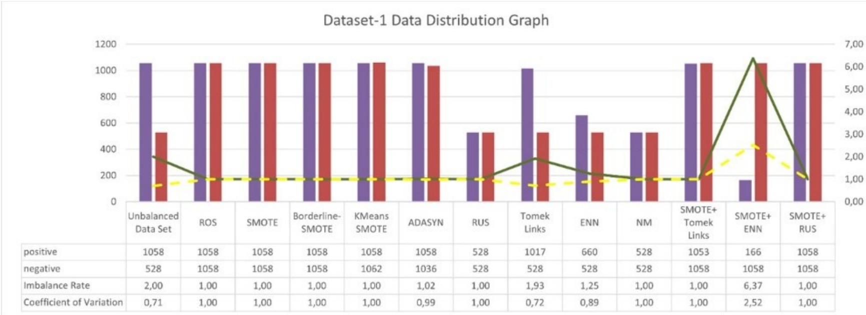


Figure 1: Dataset1 data distributions

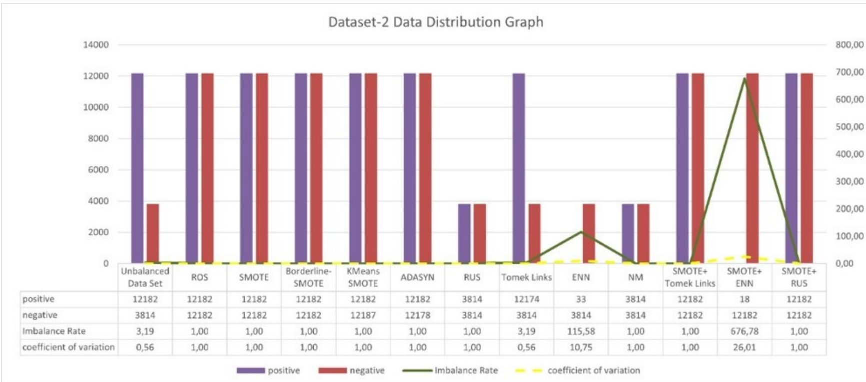
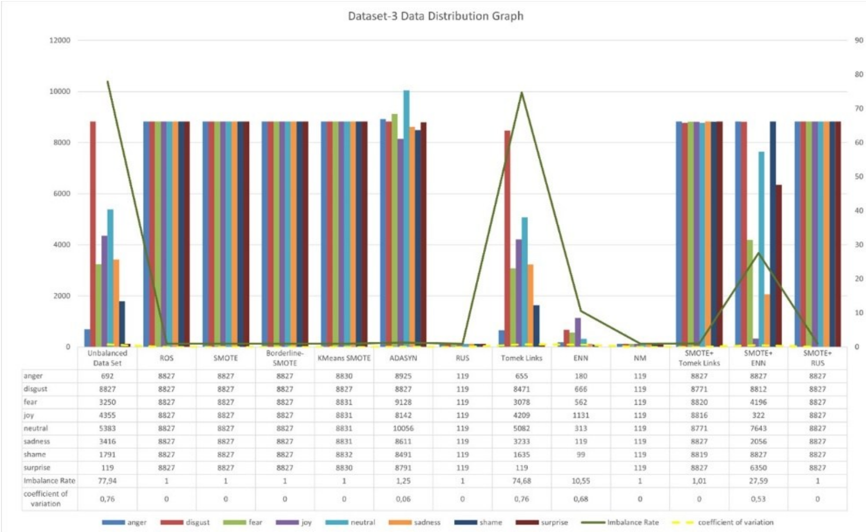


Figure 2: Dataset2 data distributions

As seen in Figures 1 and 2, the imbalance ratio for Dataset-1 was calculated as 2.00 with a coefficient of variance of 0.71, while for Dataset-2, the imbalance ratio was 3.19 with a coefficient of variance of 0.56. For both datasets, the ROS, SMOTE, Borderline-SMOTE, KMeans-SMOTE, ADASYN, SMOTE+Tomek Links, and SMOTE+RUS methods increased the number of positive examples, thus approaching or equalizing the imbalance ratio and coefficient of variance to around 1.00 by either increasing or balancing the positive examples. Conversely, RUS and NM methods reduced the number of positive examples. Tomek Links, ENN, and SMOTE+ENN methods did not achieve a balance between classes.



	Unbalanced Data Set	ROS	SMOTE	Borderline-SMOTE	KMeans SMOTE	ADASYN	RUS	Tomek Links	ENN	NM	SMOTE+ Tomek Links	SMOTE+ ENN	SMOTE+ RUS
anger	692	8827	8827	8827	8830	8925	119	655	180	119	8827	8827	8827
disgust	8827	8827	8827	8827	8827	8827	119	8471	666	119	8771	8812	8827
fear	3250	8827	8827	8827	8831	9128	119	3078	562	119	8820	4196	8827
joy	4355	8827	8827	8827	8831	8142	119	4209	1131	119	8816	322	8827
neutral	5383	8827	8827	8827	8831	10056	119	5082	313	119	8771	7643	8827
sadness	3416	8827	8827	8827	8831	8611	119	3233	119	119	8827	2056	8827
shame	1791	8827	8827	8827	8832	8491	119	1635	99	119	8819	8827	8827
surprise	119	8827	8827	8827	8830	8791	119	119		119	8827	6350	8827
Imbalance Rate	77,94	1	1	1	1	1,25	1	74,68	10,55	1	1,01	27,59	1
coefficient of variation	0,76	0	0	0	0	0,06	0	0,76	0,68	0	0	0,53	0

Figure 3: Dataset3 data distributions

As seen in Figure 3, for the multi-class Dataset-3, the initial imbalance ratio was calculated as 77.94 and the coefficient of variance as 0.76. Methods such as ROS, SMOTE, Borderline-SMOTE, KMeans-SMOTE, ADASYN, SMOTE+Tomek Links, and SMOTE+RUS increase the number of examples in classes. On the other hand, RUS and NM methods reduce the number of samples in classes, thereby equalizing the imbalance ratio and variance coefficient to 1.00 or very close to it. Tomek Links, ENN, and SMOTE+ENN methods, unlike the other two datasets, could not achieve a balanced distribution between classes.

In conclusion, across all three datasets, ROS, SMOTE, Borderline-SMOTE, KMeans-SMOTE, ADASYN, RUS, NM, SMOTE+Tomek Links, and SMOTE+RUS methods successfully address the imbalance problem in datasets and balance the coefficient of variance. However, Tomek Links, ENN, and SMOTE+ENN methods were insufficient in achieving inter-class balance in the datasets.

2. Classification Results

In the study, imbalanced datasets and balanced datasets using various methods were provided to NB, LR, SVM, DT, and RF classifiers. Their performances were compared using metrics such as accuracy, precision, recall, f1-score, and g-mean. Tables 5a, 5b, 6a, 6b, 7a and 7b respectively present the classification performances obtained from Dataset-1, Dataset-2, and Dataset-3.

Additionally, ROC curves, which visualize the relationship between TPR (True Positive Rate) and FPR (False Positive Rate) at varying threshold values, and the AUC (Area Under the Curve) value, which provides information about the overall performance of the model, are presented in Figures 4, 5, and 6.

Table 5a: Classification Performance with Balancing Methods on Dataset1 (Part 1)

		Acc	Prec	Rec	F1	GM
imbalanced dataset	**NB**	0,7406	0,7235	0,9848	0,8341	0,5072
	LR	0,7531	0,7378	0,9734	0,8393	0,5589
	SVM	0,7935	0,8027	0,9125	0,8541	0,7147
	DT	0,7305	0,7826	0,8213	0,8015	0,6735
	RF	0,7733	0,7582	0,9658	0,8495	0,6181
dataset balanced with **ROS**	**NB**	0,8212	0,8780	0,8479	0,8627	0,8073
	LR	0,8035	0,8339	0,8783	0,8556	0,7595
	SVM	0,8161	0,8345	0,9011	0,8665	0,7649
	DT	0,7078	0,7976	0,7490	0,7725	0,6852
	RF	0,7834	0,8242	0,8555	0,8396	0,7410
dataset balanced with **SMOTE**	**NB**	0,8237	0,8697	0,8631	0,8664	0,8026
	LR	0,8161	0,8417	0,8897	0,8651	0,7730
	SVM	0,7960	0,8273	0,8745	0,8503	0,7492
	DT	0,7229	0,7611	0,8479	0,8022	0,6364
	RF	0,7834	0,7774	0,9430	0,8522	0,6658
dataset balanced with **Borderline-SMOTE**	**NB**	0,8136	0,8706	0,8441	0,8571	0,7976
	LR	0,8086	0,8375	0,8821	0,8593	0,7654
	SVM	0,8060	0,8370	0,8783	0,8571	0,7638
	DT	0,7103	0,7868	0,7719	0,7793	0,6746
	RF	0,7758	0,7736	0,9354	0,8468	0,6579
dataset balanced with **KMeans SMOTE**	**NB**	0,8186	0,8631	0,8631	0,8631	0,7945
	LR	0,8086	0,8327	0,8897	0,8603	0,7600
	SVM	0,8035	0,8246	0,8935	0,8577	0,7484
	DT	0,7003	0,7517	0,8175	0,7832	0,6200
	RF	0,7708	0,7606	0,9544	0,8465	0,6259
dataset balanced with **ADASYN**	**NB**	0,8237	0,8755	0,8555	0,8654	0,8070
	LR	0,8136	0,8513	0,8707	0,8609	0,7815
	SVM	0,8086	0,8425	0,8745	0,8582	0,7706
	DT	0,7229	0,7931	0,7871	0,7901	0,6855
	RF	0,7834	0,7792	0,9392	0,8517	0,6697
dataset balanced with **RUS**	**NB**	0,7985	0,8765	0,8099	0,8419	0,7928
	LR	0,7809	0,8697	0,7871	0,8263	0,7778
	SVM	0,6322	0,7970	0,5970	0,6826	0,6471
	DT	0,6322	0,7970	0,5970	0,6826	0,6471
	RF	0,7229	0,9091	0,6464	0,7556	0,7513

Table 5b: Classification Performance with Balancing Methods on Dataset1 (Part 2)

		Acc	Prec	Rec	F1	GM
dataset balanced with **Tomek Links**	**NB**	0,7481	0,7296	0,9848	0,8382	0,5285
	LR	0,7582	0,7449	0,9658	0,8411	0,5820
	SVM	0,8010	0,8172	0,9011	0,8571	0,7381
	DT	0,7355	0,7821	0,8327	0,8066	0,6735
	RF	0,7783	0,7726	0,9430	0,8493	0,6552
dataset balanced with **ENN**	**NB**	0,8161	0,8442	0,8859	0,8646	0,7757
	LR	0,8035	0,8715	0,8251	0,8477	0,7925
	SVM	0,7834	0,9078	0,7490	0,8208	0,7983
	DT	0,7305	0,8645	0,7034	0,7757	0,7424
	RF	0,7557	0,9029	0,7072	0,7932	0,7757
dataset balanced with **NMU**	**NB**	0,8060	0,8523	0,8555	0,8539	0,7788
	LR	0,7935	0,8724	0,8061	0,8379	0,7871
	SVM	0,7884	0,8745	0,7947	0,8327	0,7853
	DT	0,5743	0,7176	0,5894	0,6472	0,5666
	RF	0,6499	0,8229	0,6008	0,6945	0,6696
dataset balanced with **SMOTE + Tomek Links**	**NB**	0,8262	0,8674	0,8707	0,8691	0,8021
	LR	0,7985	0,8352	0,8669	0,8507	0,7588
	SVM	0,8060	0,8345	0,8821	0,8577	0,7611
	DT	0,7204	0,7923	0,7833	0,7878	0,6838
	RF	0,7859	0,7764	0,9506	0,8547	0,6632
dataset balanced with **SMOTE + ENN**	**NB**	0,4106	1,0000	0,1103	0,1986	0,3321
	LR	0,3778	1,0000	0,0608	0,1147	0,2467
	SVM	0,4786	1,0000	0,2129	0,3511	0,4614
	DT	0,4912	0,9178	0,2548	0,3988	0,4933
	RF	0,4685	0,9815	0,2015	0,3344	0,4472
dataset balanced with **SMOTE+ RUS**	**NB**	0,8237	0,8755	0,8555	0,8654	0,8070
	LR	0,7985	0,8327	0,8707	0,8513	0,7562
	SVM	0,8060	0,8321	0,8859	0,8582	0,7584
	DT	0,7481	0,7690	0,8859	0,8233	0,6505
	RF	0,7884	0,7859	0,9354	0,8542	0,6839

Table 6a: Classification Performance with Balancing Methods on Dataset2 (Part 1)

		Acc	Prec	Rec	F1	GM
imbalanced dataset	**NB**	0,8633	0,8553	0,9882	0,9170	0,6733
	LR	0,8948	0,8984	0,9722	0,9338	0,7913
	SVM	0,9028	0,9171	0,9594	0,9378	0,8307
	DT	0,8058	0,8784	0,8655	0,8719	0,7280
	RF	0,8735	0,8795	0,9670	0,9211	0,7430
dataset balanced with **ROS**	**NB**	0,8713	0,9488	0,8789	0,9125	0,8625
	LR	0,8835	0,9484	0,8963	0,9216	0,8688
	SVM	0,8830	0,9438	0,9005	0,9216	0,8626
	DT	0,8108	0,8836	0,8665	0,8749	0,7390
	RF	0,8780	0,9134	0,9283	0,9208	0,8147
dataset balanced with **SMOTE**	**NB**	0,8733	0,9499	0,8806	0,9139	0,8649
	LR	0,8858	0,9455	0,9025	0,9235	0,8663
	SVM	0,8873	0,9381	0,9126	0,9252	0,8572
	DT	0,7918	0,8696	0,8557	0,8626	0,7074
	RF	0,8708	0,8876	0,9512	0,9183	0,7619
dataset balanced with **Borderline-SMOTE**	**NB**	0,8588	0,9524	0,8580	0,9027	0,8596
	LR	0,8708	0,9434	0,8838	0,9127	0,8557
	SVM	0,8758	0,9383	0,8963	0,9168	0,8517
	DT	0,7988	0,8740	0,8606	0,8673	0,7177
	RF	0,8545	0,8682	0,9545	0,9093	0,7117
dataset balanced with **KMeans SMOTE**	**NB**	0,8750	0,9497	0,8832	0,9152	0,8657
	LR	0,8870	0,9474	0,9022	0,9242	0,8694
	SVM	0,8863	0,9416	0,9074	0,9242	0,8614
	DT	0,7985	0,8748	0,8593	0,8670	0,7190
	RF	0,8758	0,8982	0,9444	0,9207	0,7856
dataset balanced with **ADASYN**	**NB**	0,8695	0,9522	0,8730	0,9109	0,8655
	LR	0,8755	0,9444	0,8894	0,9161	0,8594
	SVM	0,8825	0,9422	0,9015	0,9214	0,8603
	DT	0,7973	0,8706	0,8629	0,8667	0,7103
	RF	0,8668	0,8822	0,9529	0,9162	0,7485
dataset balanced with RUS	**NB**	0,8600	0,9571	0,8550	0,9032	0,8655
	LR	0,8725	0,9563	0,8730	0,9128	0,8719
	SVM	0,8663	0,9546	0,8662	0,9082	0,8663
	DT	0,7570	0,9161	0,7507	0,8252	0,7640
	RF	0,8433	0,9493	0,8397	0,8911	0,8472

Table 6b: Classification Performance with Balancing Methods on Dataset2 (Part 2)

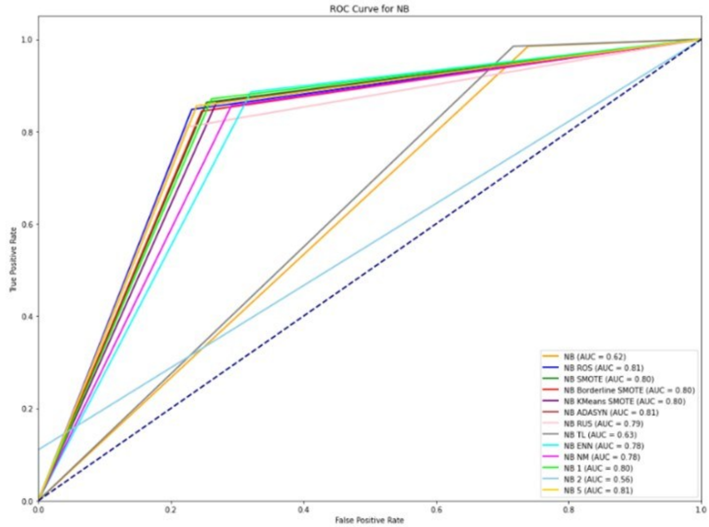
		Acc	Prec	Rec	F1	GM
dataset balanced with **Tomek Links**	**NB**	0,8633	0,8551	0,9885	0,9170	0,6726
	LR	0,8945	0,8988	0,9712	0,9336	0,7922
	SVM	0,9015	0,9162	0,9588	0,9370	0,8286
	DT	0,8088	0,8784	0,8701	0,8742	0,7286
	RF	0,8723	0,8761	0,9699	0,9206	0,7344
dataset balanced with **ENN**	**NB**	0,2360	0,0000	0,0000	0,0000	0,0000
	LR	0,2360	0,0000	0,0000	0,0000	0,0000
	SVM	0,2415	1,0000	0,0072	0,0143	0,0848
	DT	0,2455	0,8958	0,0141	0,0277	0,1183
	RF	0,2388	1,0000	0,0036	0,0072	0,0600
dataset balanced with **NMU**	**NB**	0,8610	0,9545	0,8590	0,9042	0,8633
	LR	0,8760	0,9581	0,8760	0,9152	0,8760
	SVM	0,8708	0,9549	0,8721	0,9116	0,8693
	DT	0,7520	0,8994	0,7605	0,8241	0,7423
	RF	0,8425	0,9499	0,8380	0,8905	0,8475
dataset balanced with **SMOTE + Tomek**	**NB**	0,8723	0,9485	0,8806	0,9133	0,8628
	LR	0,8850	0,9430	0,9041	0,9232	0,8627
	SVM	0,8910	0,9417	0,9139	0,9276	0,8640
	DT	0,7928	0,8713	0,8550	0,8631	0,7109
	RF	0,8685	0,8852	0,9512	0,9170	0,7559
dataset balanced with **SMOTE + ENN**	**NB**	0,2360	0,0000	0,0000	0,0000	0,0000
	LR	0,2360	0,0000	0,0000	0,0000	0,0000
	SVM	0,2365	1,0000	0,0007	0,0013	0,0256
	DT	0,2393	0,8824	0,0049	0,0098	0,0700
	RF	0,2360	0,0000	0,0000	0,0000	0,0000
dataset balanced with **SMOTE+ RUS**	**NB**	0,8723	0,9479	0,8812	0,9133	0,8620
	LR	0,8840	0,9435	0,9022	0,9224	0,8628
	SVM	0,8878	0,9399	0,9113	0,9254	0,8599
	DT	0,7958	0,8740	0,8560	0,8649	0,7170
	RF	0,8650	0,8854	0,9457	0,9146	0,7557

Table 7a: Classification Performance with Balancing Methods on Dataset3 (Part 1)

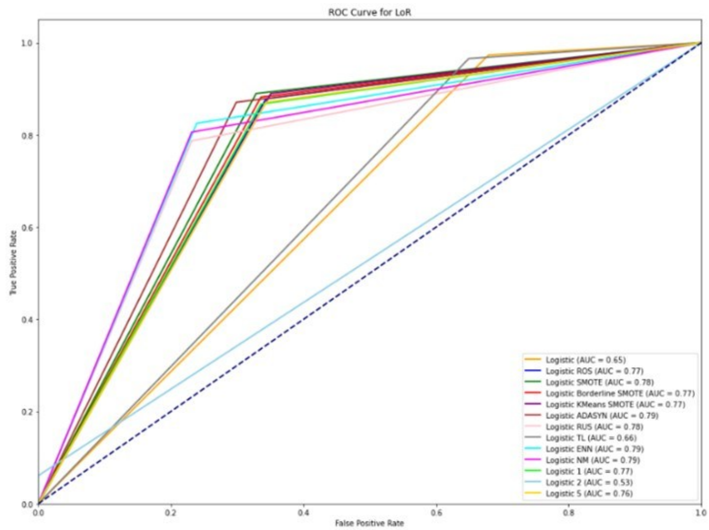
		Acc	Prec	Rec	F1	GM
imbalanced dataset	**NB**	0,5540	0,5951	0,5540	0,5239	0,2939
	LR	0,6055	0,6185	0,6055	0,5922	0,4073
	SVM	0,6178	0,6266	0,6178	0,6076	0,5098
	DT	0,5478	0,5449	0,5478	0,5434	0,4561
	RF	0,6191	0,6420	0,6191	0,6057	0,5053
dataset balanced with **ROS**	**NB**	0,5492	0,5905	0,5492	0,5632	0,4129
	LR	0,5942	0,6196	0,5942	0,6010	0,5287
	SVM	0,5964	0,6095	0,5964	0,5994	0,5231
	DT	0,5360	0,5399	0,5360	0,5358	0,4768
	RF	0,5985	0,6091	0,5985	0,5908	0,5109
dataset balanced with **SMOTE**	**NB**	0,5522	0,5888	0,5522	0,5646	0,4199
	LR	0,5997	0,6185	0,5997	0,6056	0,5316
	SVM	0,5949	0,6063	0,5949	0,5982	0,5225
	DT	0,5302	0,5312	0,5302	0,5289	0,4719
	RF	0,5890	0,6058	0,5890	0,5847	0,5029
dataset balanced with **Borderline -SMOTE**	**NB**	0,5453	0,5490	0,5453	0,5448	0,3739
	LR	0,5897	0,5896	0,5897	0,5876	0,4724
	SVM	0,5919	0,5927	0,5919	0,5884	0,5001
	DT	0,5298	0,5277	0,5298	0,5257	0,4337
	RF	0,5823	0,6079	0,5823	0,5743	0,4719
dataset balanced with **KMeans SMOTE**	**NB**	0,5465	0,5680	0,5465	0,5544	0,3887
	LR	0,5800	0,5912	0,5800	0,5835	0,4894
	SVM	0,5744	0,5835	0,5744	0,5759	0,4875
	DT	0,5405	0,5403	0,5405	0,5389	0,4735
	RF	0,6153	0,6215	0,6153	0,6073	0,5238
dataset balanced with **ADASYN**	**NB**	0,5435	0,5944	0,5435	0,5590	0,4158
	LR	0,5926	0,6147	0,5926	0,5997	0,5261
	SVM	0,5915	0,6070	0,5915	0,5962	0,5181
	DT	0,5274	0,5301	0,5274	0,5259	0,4644
	RF	0,5777	0,6028	0,5777	0,5771	0,4977
dataset balanced with **RUS**	**NB**	0,3383	0,4156	0,3383	0,3590	0,2584
	LR	0,3591	0,4296	0,3591	0,3745	0,3564
	SVM	0,3631	0,4219	0,3631	0,3758	0,3627
	DT	0,2737	0,3932	0,2737	0,2903	0,2576
	RF	0,2677	0,4727	0,2677	0,2954	0,2611

Table 7b: Classification Performance with Balancing Methods on Dataset3 (Part 2)

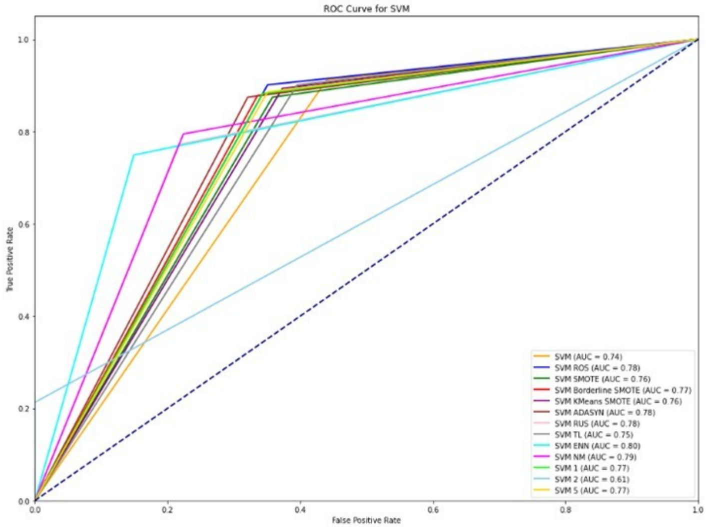
		Acc	Prec	Rec	F1	GM
dataset balanced with **Tomek Links**	**NB**	0,5553	0,5998	0,5553	0,5244	0,2930
	LR	0,6037	0,6173	0,6037	0,5901	0,4055
	SVM	0,6199	0,6289	0,6199	0,6101	0,5122
	DT	0,5518	0,5499	0,5518	0,5479	0,4677
	RF	0,6192	0,6423	0,6192	0,6064	0,5057
dataset balanced with **ENN**	**NB**	0,2413	0,6988	0,2413	0,2555	0,1341
	LR	0,2196	0,7618	0,2196	0,2504	0,1999
	SVM	0,2595	0,7196	0,2595	0,3072	0,2639
	DT	0,3340	0,5140	0,3340	0,3602	0,2825
	RF	0,3174	0,6455	0,3174	0,3594	0,2922
dataset balanced with **NMU**	**NB**	0,2644	0,3810	0,2644	0,2863	0,2196
	LR	0,1796	0,4884	0,1796	0,1982	0,2524
	SVM	0,1328	0,4875	0,1328	0,1308	0,2148
	DT	0,2065	0,4062	0,2065	0,2233	0,2153
	RF	0,2039	0,4819	0,2039	0,2244	0,2583
dataset balanced with **SMOTE + Tomek**	**NB**	0,5532	0,5917	0,5532	0,5663	0,4227
	LR	0,5994	0,6184	0,5994	0,6054	0,5321
	SVM	0,5939	0,6059	0,5939	0,5972	0,5202
	DT	0,5302	0,5305	0,5302	0,5285	0,4706
	RF	0,5932	0,6071	0,5932	0,5871	0,5064
dataset balanced with **SMOTE + ENN**	**NB**	0,2871	0,6789	0,2871	0,2339	0,2641
	LR	0,3677	0,6628	0,3677	0,3160	0,4022
	SVM	0,3763	0,6115	0,3763	0,3326	0,4010
	DT	0,3545	0,5424	0,3545	0,3224	0,3823
	RF	0,3617	0,6209	0,3617	0,3194	0,3849
dataset balanced with **SMOTE+ RUS**	**NB**	0,5481	0,5871	0,5481	0,5613	0,4176
	LR	0,5982	0,6166	0,5982	0,6039	0,5315
	SVM	0,5926	0,6040	0,5926	0,5957	0,5218
	DT	0,5341	0,5355	0,5341	0,5324	0,4739
	RF	0,5920	0,6098	0,5920	0,5872	0,5030


ROC Curve for NB
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
NB (AUC = 0.62)
NB ROS (AUC = 0.81)
NB SMOTE (AUC = 0.80)
NB Borderline SMOTE (AUC = 0.80)
NB KMeans SMOTE (AUC = 0.80)
NB ADASYN (AUC = 0.81)
NB RUS (AUC = 0.79)
NB TL (AUC = 0.63)
NB ENN (AUC = 0.78)
NB NM (AUC = 0.78)
NB 1 (AUC = 0.80)
NB 2 (AUC = 0.56)
NB 5 (AUC = 0.81)

a. NB

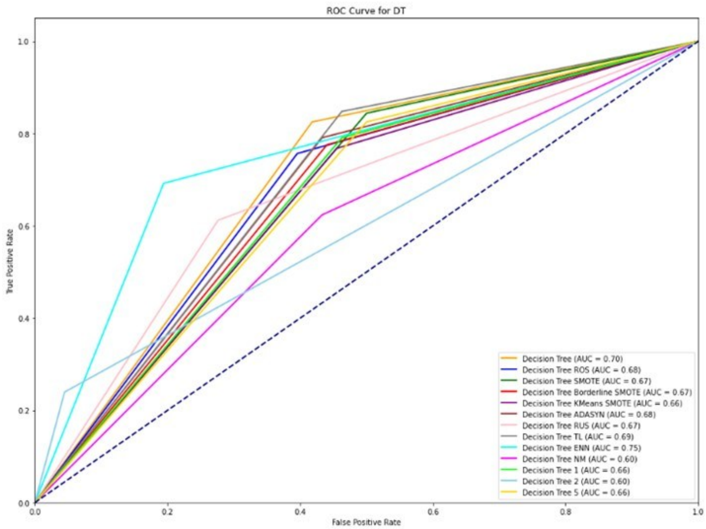

ROC Curve for LoR
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic (AUC = 0.65)
Logistic ROS (AUC = 0.77)
Logistic SMOTE (AUC = 0.78)
Logistic Borderline SMOTE (AUC = 0.77)
Logistic KMeans SMOTE (AUC = 0.77)
Logistic ADASYN (AUC = 0.79)
Logistic RUS (AUC = 0.78)
Logistic TL (AUC = 0.66)
Logistic ENN (AUC = 0.79)
Logistic NM (AUC = 0.79)
Logistic 1 (AUC = 0.77)
Logistic 2 (AUC = 0.53)
Logistic 5 (AUC = 0.76)

b. LR



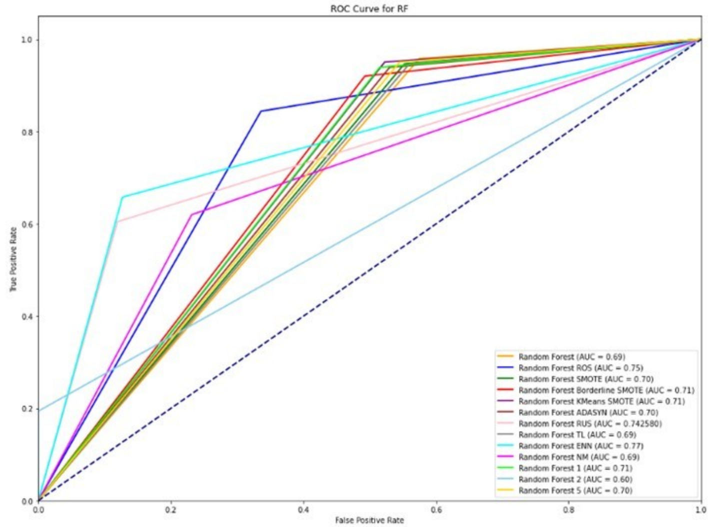
ROC Curve for SVM
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
SVM (AUC = 0.74)
SVM ROS (AUC = 0.78)
SVM SMOTE (AUC = 0.76)
SVM Borderline SMOTE (AUC = 0.77)
SVM KMeans SMOTE (AUC = 0.76)
SVM ADASYN (AUC = 0.78)
SVM RUS (AUC = 0.78)
SVM TL (AUC = 0.75)
SVM ENN (AUC = 0.80)
SVM NM (AUC = 0.79)
SVM 1 (AUC = 0.77)
SVM 2 (AUC = 0.61)
SVM 5 (AUC = 0.77)

c. SVM



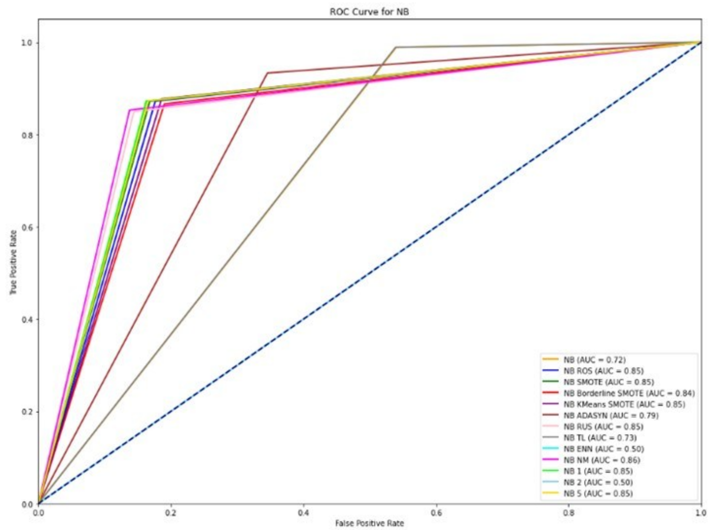
ROC Curve for DT
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Decision Tree (AUC = 0.70)
Decision Tree ROS (AUC = 0.68)
Decision Tree SMOTE (AUC = 0.67)
Decision Tree Borderline SMOTE (AUC = 0.67)
Decision Tree KMeans SMOTE (AUC = 0.66)
Decision Tree ADASYN (AUC = 0.68)
Decision Tree RUS (AUC = 0.67)
Decision Tree TL (AUC = 0.69)
Decision Tree ENN (AUC = 0.75)
Decision Tree NM (AUC = 0.60)
Decision Tree 1 (AUC = 0.66)
Decision Tree 2 (AUC = 0.60)
Decision Tree 5 (AUC = 0.66)

d. DT

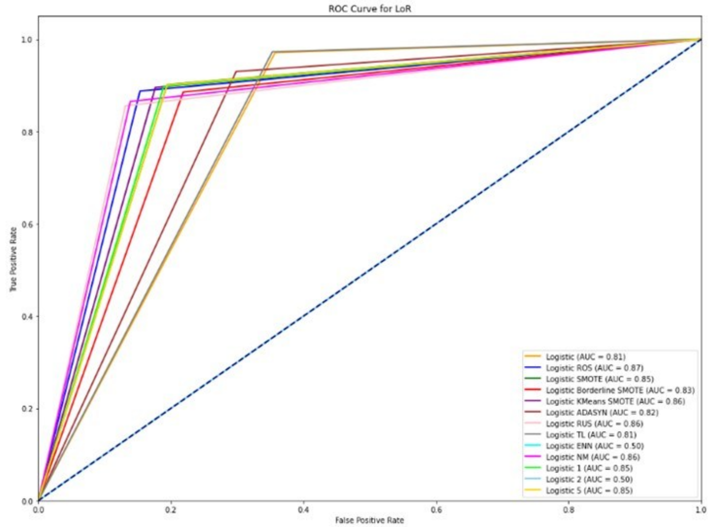


e. RF

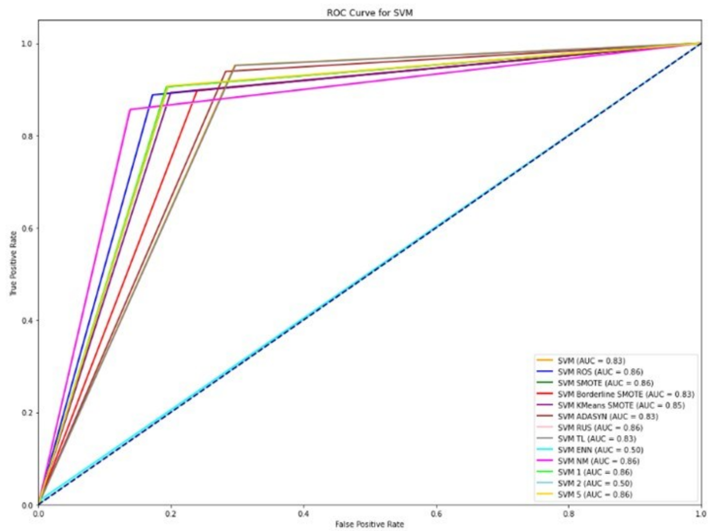
Figure 4. ROC curve and AUC values of classifiers for Dataset1



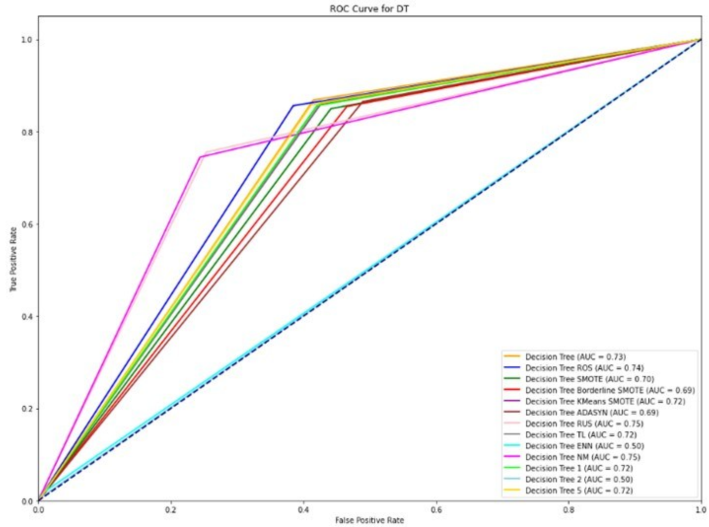
a. NB


ROC Curve for LoR
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic (AUC = 0.81)
Logistic ROS (AUC = 0.87)
Logistic SMOTE (AUC = 0.85)
Logistic Borderline SMOTE (AUC = 0.83)
Logistic KMeans SMOTE (AUC = 0.86)
Logistic ADASYN (AUC = 0.82)
Logistic RUS (AUC = 0.86)
Logistic TL (AUC = 0.81)
Logistic ENN (AUC = 0.50)
Logistic NM (AUC = 0.86)
Logistic 1 (AUC = 0.85)
Logistic 2 (AUC = 0.50)
Logistic 5 (AUC = 0.85)

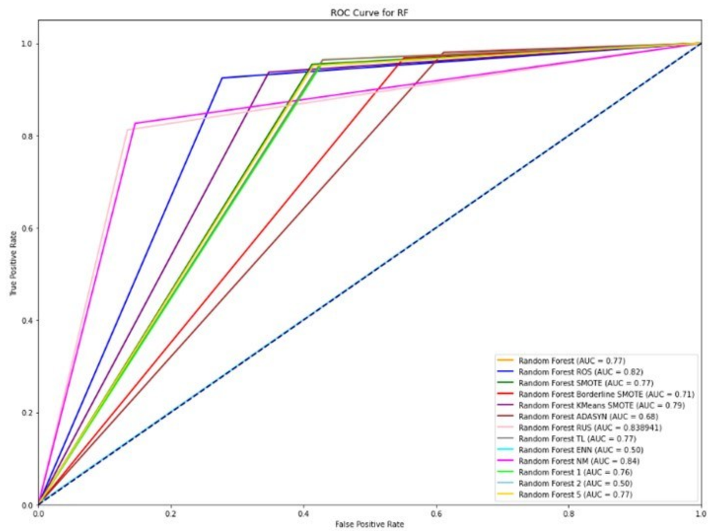
b. LR


ROC Curve for SVM
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
SVM (AUC = 0.83)
SVM ROS (AUC = 0.86)
SVM SMOTE (AUC = 0.86)
SVM Borderline SMOTE (AUC = 0.83)
SVM KMeans SMOTE (AUC = 0.85)
SVM ADASYN (AUC = 0.83)
SVM RUS (AUC = 0.86)
SVM TL (AUC = 0.83)
SVM ENN (AUC = 0.50)
SVM NM (AUC = 0.86)
SVM 1 (AUC = 0.86)
SVM 2 (AUC = 0.50)
SVM 5 (AUC = 0.86)

c. SVM

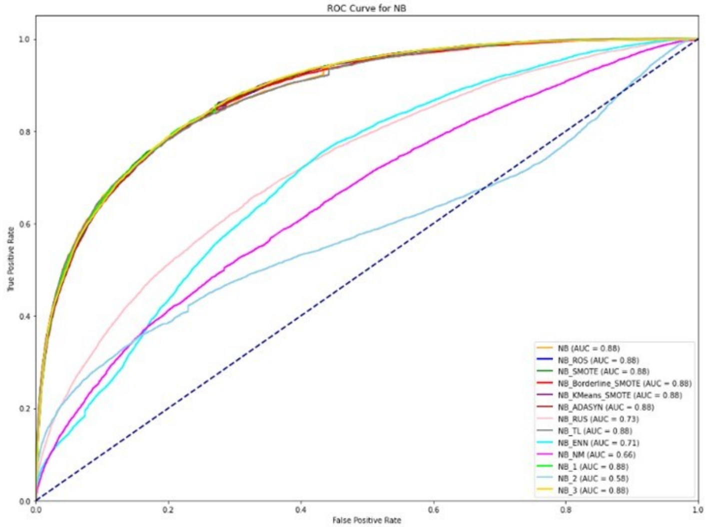


d. DT

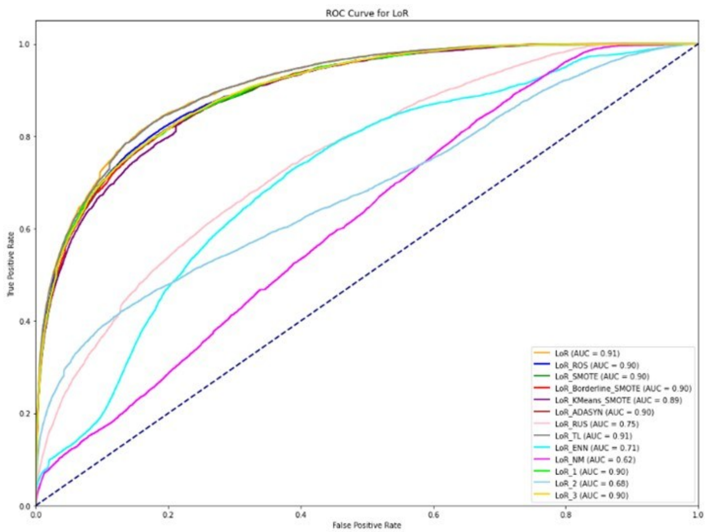


e. RF

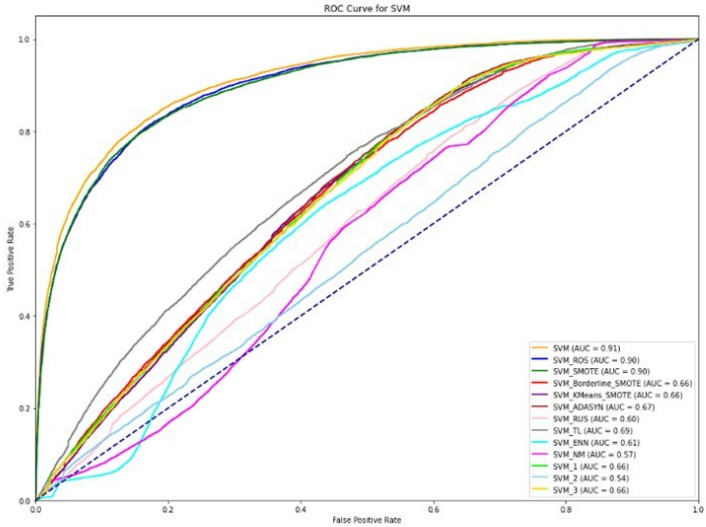
Figure 5. ROC curve and AUC values of classifiers for Dataset2


ROC Curve for NB
True Positive Rate
False Positive Rate
NB (AUC = 0.88)
NB_ROS (AUC = 0.88)
NB_SMOTE (AUC = 0.88)
NB_Borderline_SMOTE (AUC = 0.88)
NB_KMeans_SMOTE (AUC = 0.88)
NB_ADASYN (AUC = 0.88)
NB_RUS (AUC = 0.73)
NB_TL (AUC = 0.88)
NB_ENN (AUC = 0.71)
NB_NM (AUC = 0.66)
NB_1 (AUC = 0.88)
NB_2 (AUC = 0.58)
NB_3 (AUC = 0.88)

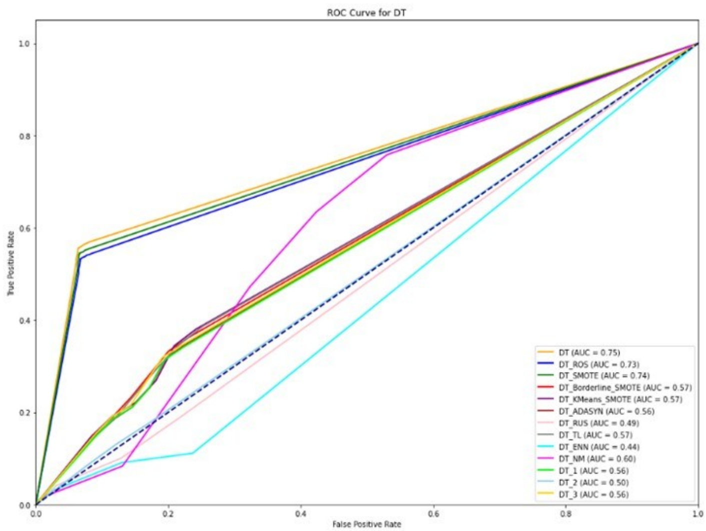
a. NB


ROC Curve for LoR
True Positive Rate
False Positive Rate
LoR (AUC = 0.91)
LoR_ROS (AUC = 0.90)
LoR_SMOTE (AUC = 0.90)
LoR_Borderline_SMOTE (AUC = 0.90)
LoR_KMeans_SMOTE (AUC = 0.89)
LoR_ADASYN (AUC = 0.90)
LoR_RUS (AUC = 0.75)
LoR_TL (AUC = 0.91)
LoR_ENN (AUC = 0.71)
LoR_NM (AUC = 0.62)
LoR_1 (AUC = 0.90)
LoR_2 (AUC = 0.68)
LoR_3 (AUC = 0.90)

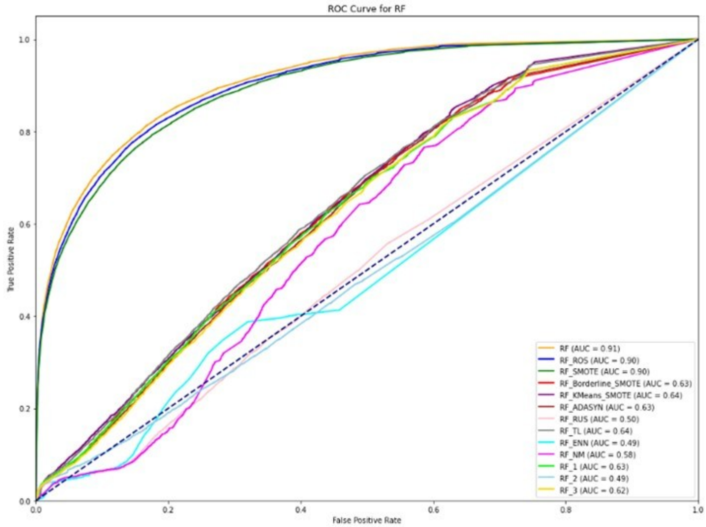
b. LR


ROC Curve for SVM
True Positive Rate
False Positive Rate
SVM (AUC = 0.91)
SVM_ROS (AUC = 0.90)
SVM_SMOTE (AUC = 0.90)
SVM_Borderline_SMOTE (AUC = 0.66)
SVM_KMeans_SMOTE (AUC = 0.66)
SVM_ADASYN (AUC = 0.67)
SVM_RUS (AUC = 0.60)
SVM_TL (AUC = 0.69)
SVM_ENN (AUC = 0.61)
SVM_NM (AUC = 0.57)
SVM_1 (AUC = 0.66)
SVM_2 (AUC = 0.54)
SVM_3 (AUC = 0.66)

c. SVM


ROC Curve for DT
True Positive Rate
False Positive Rate
DT (AUC = 0.75)
DT_ROS (AUC = 0.73)
DT_SMOTE (AUC = 0.74)
DT_Borderline_SMOTE (AUC = 0.57)
DT_KMeans_SMOTE (AUC = 0.57)
DT_ADASYN (AUC = 0.56)
DT_RUS (AUC = 0.49)
DT_TL (AUC = 0.57)
DT_ENN (AUC = 0.44)
DT_NM (AUC = 0.60)
DT_1 (AUC = 0.56)
DT_2 (AUC = 0.50)
DT_3 (AUC = 0.56)

d. DT



e. RF

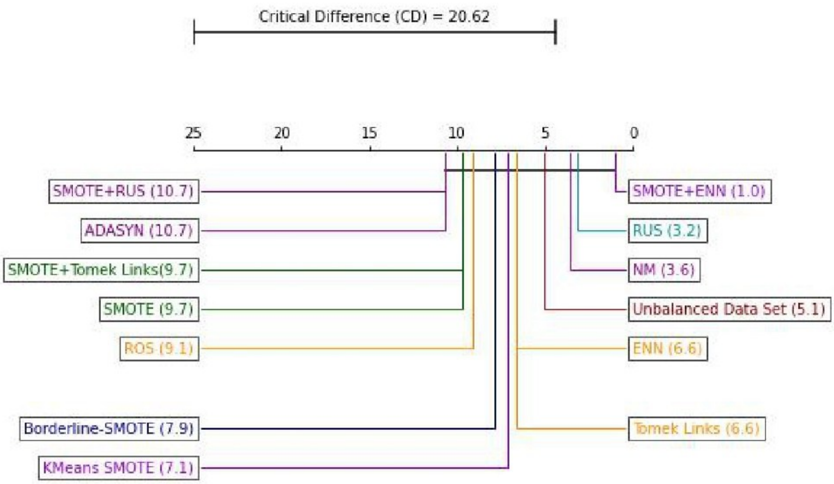
Figure 6. ROC curve and AUC values of classifiers for Dataset3

When compared using the experimental results provided in Tables 5a-5b-6a-6b-7a-7b and Figures 4-5-6, the following conclusions were obtained regarding the performance of various balancing methods and classifier models used to enhance classification performance:

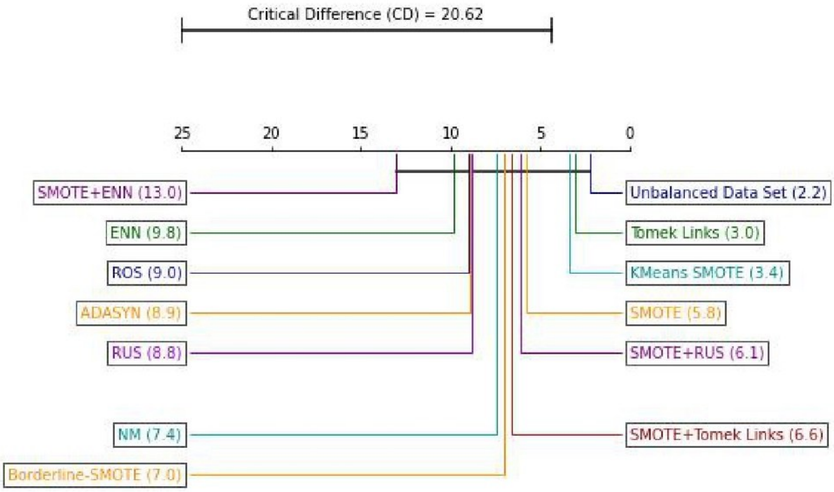
- Before applying balancing methods, SVM generally outperformed other classifiers when the three datasets were provided to the models.
- After applying various balancing methods, the NB algorithm generally achieved the best performance. There was not a significant improvement observed in the performance of LR, SVM, and DT algorithms with balancing methods.
- Over-sampling methods such as SMOTE, Borderline-SMOTE, KMeans-SMOTE, and the hybrid method SMOTE+Tomek Links effectively balanced the imbalances in the datasets and improved performance.

- ENN and NMU methods generally did not improve performance and in some cases led to a decrease in performance.

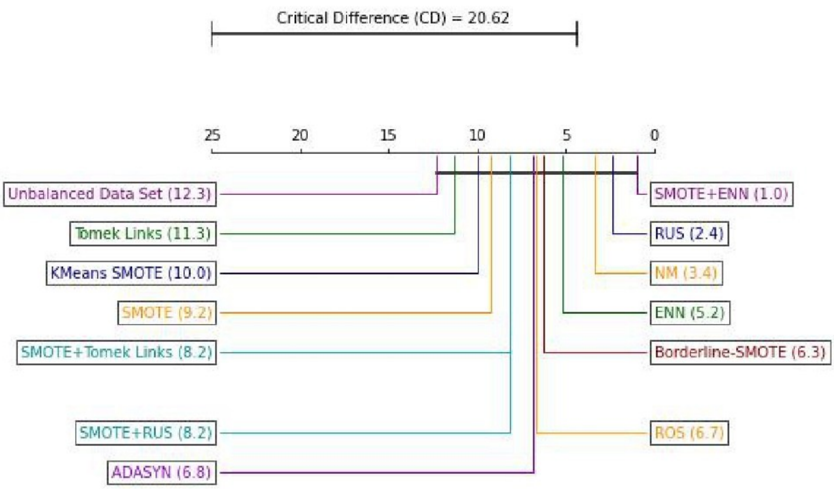
Additionally, after applying balancing methods to Dataset-1, statistical analysis using Friedman and Nemenyi tests was conducted to compare the performance values obtained from various classifiers. Critical diagrams generated by the Nemenyi post-hoc test for accuracy, precision, recall, and F1-score metrics are presented in Figure 7.



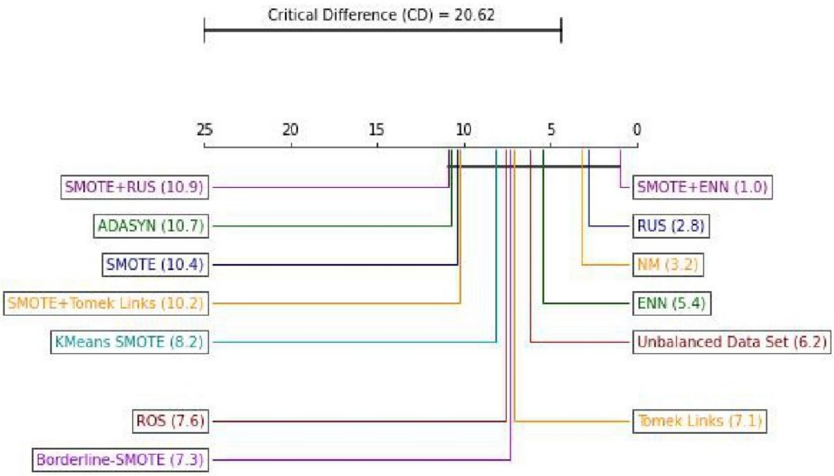
a. accuracy



b. precision



c. recall



d. F score

Figure 7. Nemenyi critical diagram for Dataset1

First, the average order of the data balancing methods applied in the Friedman and Nemenyi tests was calculated. Test statistics for accuracy, precision, recall and f1-score calculated using the Friedman test were 37.64, 36.04, 45.56 and 40.88, respectively. When the α value was taken as 0.05 for a total of five architectures: NB, LR, SVM, DT and RF, the critical test statistic was obtained as 21.026. In order to refute the null hypothesis test, the test statistics values obtained by the Friedman test must be higher than the critical test statistic. When the values were compared, the inequality of 37.64>21.026 for classification accuracy, 36.04>21.026 for sensitivity, 45.56>21.026 for recall and 40.88>21.026 for the f1 criterion was achieved, so the null hypothesis test was rejected and the post-hoc test was applied.

$$CD = q * \sqrt{}(k * (k + 1)/(6.0 * N)) \tag{8}$$

To determine the critical difference (CD) value for the Nemenyi test, Equation 8 was used. In this formula, q represents a specific critical value corresponding to a given confidence level, calculated using the Studentized Range q table (Zaiontz). Here, k denotes the number of models evaluated in the system, and N represents the number of data points within each group. For the study conducted on Dataset-1, k=13 and N=5 were used. The CD value was computed as 20.62 using Equation 8.

According to the average results obtained from the Nemenyi test, models created using datasets balanced with KMeans-SMOTE for accuracy metric, Borderline-SMOTE for precision and F1-score metrics, and ADASYN for recall metric have shown better performance compared to models created using other balancing methods. These models have demonstrated results closest to the averages of models created with other balancing methods, indicating a relationship among the established models overall.

Conclusion

In this study, a specific dataset was created by collecting tweets tagged with "Metaverse" from the X platform; in addition, sentiment analysis was conducted using two different pre-existing datasets. The analysis investigated the performance effects of 12 different data balancing methods on traditional machine learning classifiers.

The experimental results have shown that SVM classifier achieved the highest performance on imbalanced datasets, while NB classifier performed best on balanced datasets. Among the data balancing methods used, techniques like SMOTE, Borderline-SMOTE, KMeans-SMOTE oversampling methods, and the SMOTE+Tomek Links hybrid method generally improved classifier performance. Furthermore, the performance results from data balancing methods were subjected to the Nemenyi statistical test, revealing significant relationships among the models built.

For future studies, comparisons are planned not only at the data preprocessing level using data balancing methods but also at the algorithmic level and with ensemble methods. Additionally, increasing the number of tweets in the dataset to work with a larger dataset is also an objective.

References

Bagui, S., Mink, D., Bagui, S., Subramaniam, S., & Wallace, D. (2023). Resampling Imbalanced Network Intrusion Datasets to Identify Rare Attacks. *Future Internet, 15*(4). doi:10.3390/fi15040130

Batista, G., Prati, R.C., & Monard, M.C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter, 6*(1), 20-29.

Brownlee, J. (2020). Undersampling algorithms for imbalanced classification. *Machine Learning Mastrey, 27*.

Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357.

Cheng, S. (2023). Metaverse. In *Metaverse: Concept, Content and Context* (pp. 1-23). Cham: Springer Nature Switzerland.

Dewi, C., Chen, R.C., Christanto, H.J., & Cauteruccio, F. (2023). Multinomial Naïve Bayes Classifier for Sentiment Analysis of Internet Movie Database. *Vietnam Journal of Computer Science (World Scientific), 10*(4).

Dina, A. S., Siddique, A. B., & Manivannan, D. (2022). Effect of Balancing Data Using Synthetic Data on the Performance of Machine Learning Classifiers for Intrusion Detection in Computer Networks. *Ieee Access, 10*, 96731-96747. doi:10.1109/access.2022.3205337

Domingues, I., Amorim, J.P., Abreu, P.H., Duarte, H., & Santos, J. (2018). *Evaluation of oversampling data balancing techniques in the context of ordinal classification.* Paper presented at the 2018 International Joint Conference on Neural Networks (IJCNN).

Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information sciences, 465*, 1-20.

Duan, H. J., Wei, Y. Q., Liu, P. Y., & Yin, H. X. (2020). A Novel Ensemble Framework Based on K-Means and Resampling for Imbalanced Data. *Applied Sciences-Basel, 10*(5). doi:10.3390/app10051684

Ebrahimy, H., Mirbagheri, B., Matkan, A.A., & Azadbakht, M. (2022). Effectiveness of the integration of data balancing techniques and tree-based ensemble machine learning algorithms for spatially-explicit land cover accuracy prediction. *Remote Sensing Applications: Society and Environment, 27*, 100785.

Góra, G., & Skowron, A. (2023). *On kNN Class Weights for Optimising G-Mean and F1-Score.* Paper presented at the International Joint Conference on Rough Sets.

Habek, G.C. (2022). *Makine Öğrenmesi Teknikleriyle Kripto Para Duygu Analizi.* (Master Thesis), Manisa Celal Bayar Üniversitesi, Manisa.

Habek, G.C. (2023). *Metaverse İle İlgili Twitter Verilerinin Makine Öğrenmesi Ve Topluluk Tabanlı Yöntemler Kullanılarak Sınıflandırılması*. Paper presented at the 10th International Congress on Life, Engineering, and Applied Sciences in a Changing World.

Halimu, C., & Kasem, A. (2021). *Split balancing (sBal)—A data preprocessing sampling technique for ensemble methods for binary classification in imbalanced datasets.* Paper presented at the

Computational Science and Technology: 7th ICCST 2020, Pattaya, Thailand, 29–30 August, 2020.

Han, H., Wang, W.Y., & Mao, B.H. (2005). *Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning.* Paper presented at the International conference on intelligent computing.

Hart, P. (1968). *The condensed nearest neighbor rule (Corresp.)*. Paper presented at the IEEE Transactions on Information Theory.

He, H., Bai, Y., Garcia, E.A., & Li, S. (2008). *ADASYN: Adaptive synthetic sampling approach for imbalanced learning.* Paper presented at the 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence).

He, H., & Garcia, E.A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering, 21*(9), 1263-1284.

Jadhav, A., Samih, M.M., Elmannai, H., & Karim, F.K. (2022). An Empirical Assessment of Performance of Data Balancing Techniques in Classification Task. *Applied Sciences, 12*(8), 3928.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review, gests international transactions on computer science and engineering 30 (2006) 25–36. *Synthetic Oversampling of Instances Using Clustering*.

Kubat, M., & Matwin, S. (1997). *Addressing the curse of imbalanced training sets: one-sided selection.* Paper presented at the Icml.

Mqadi, N.M., Naicker, N., & Adeliyi, T. (2021). Solving Misclassification of the Credit Card Imbalance Problem Using Near Miss. *Mathematical Problems in Engineering, 2021*, 7194728. doi:10.1155/2021/7194728

Muschelli, J. (2020). ROC and AUC with a Binary Predictor: a Potentially Misleading Metric. *Journal of Classification, 37*(3), 696-708. doi:10.1007/s00357-019-09345-1

Nurkholis, A., Alita, D., & Munandar, A. (2022). Comparison of Kernel Support Vector Machine Multi-Class in PPKM Sentiment Analysis on Twitter. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 6*(2), 227-233.

Potharaju, S. P., Sreedevi, M., Ande, V. K., & Tirandasu, R. K. (2019). Data mining approach for accelerating the classification accuracy of cardiotocography. *Clinical Epidemiology and Global Health, 7*(2), 160-164. doi:10.1016/j.cegh.2018.03.004

Saripuddin, M., Suliman, A., S.S., Syarmila, & Jorgensen, B.N. (2021). *Random undersampling on imbalance time series data for anomaly detection.* Paper presented at the Proceedings of the 2021 4th International Conference on Machine Learning and Machine Intelligence.

Stando, A., Cavus, M., & Biecek, P. (2024). *The effect of balancing methods on model behavior in imbalanced classification problems.* Paper presented at the Fifth International Workshop on Learning with Imbalanced Domains: Theory and Applications.

Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., & Zhou, Y. (2015). A novel ensemble method for classifying imbalanced data. *Pattern Recognition, 48*(5), 1623-1637.

Tahir, M.A., Kittler, J., Mikolajczyk, K., & Yan, F. (2009). *A multiple expert approach to the class imbalance problem using inverse random under sampling.* Paper presented at the Multiple Classifier Systems: 8th International Workshop, MCS 2009, Reykjavik, Iceland, June 10-12, 2009. Proceedings 8.

Tolba, M., Ouadfel, S., & Meshoul, S. (2021). Hybrid ensemble approaches to online harassment detection in highly imbalanced data. *Expert Systems with Applications, 175*, 114751.

Tomek, I. (1976). Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-6*, 769-772. doi:10.1109/TSMC.1976.4309452

Wang, Y., Su, Z., Zhang, N., Xing, R., Liu, D., Luan, T.H., & Shen, X. (2022). A survey on metaverse: Fundamentals, security, and privacy. *IEEE Communications Surveys & Tutorials, 25*(1), 319-352.

Wilson, D. L. . (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-2*, 408-421. doi:10.1109/TSMC.1972.4309137

Zaiontz, Charles. Studentized Range q Table. Retrieved from https://real-statistics.com/statistics-tables/studentized-range-q-table/

CHAPTER II

An Approach to Enhancing Efficiency in Image Classification with CNN Models

Mehmet Bilge Han TAŞ[1]
Eyyüp YILDIZ[2]

1. Introduction

Artificial intelligence studies In the 1950s, Alan Turing's "Can machines think?" started by seeking an answer to the question (Turing & Haugeland, 1950). Later, in Cahit Arf's valuable work, answers were sought to questions such as whether machines have the ability to think and, if so, how can thinking machines be created (Arf, 1959). Artificial intelligence studies continued with this curiosity and desire until the 1990s, but the acceleration of the studies decreased due to insufficient processing power and insufficient data. Later, with the developing technology and the widespread use of the

[1] Res. Asst., Erzincan Binali Yildirim University, Faculty of Engineering and Architecture, Department of Computer Engineering, Erzincan, Turkey, ORCID: 0000-0001-6135-1849 bilgehantas@erzincan.edu.tr
[2] Asst. Prof. Dr., Erzincan Binali Yildirim University, Faculty of Engineering and Architecture, Department of Computer Engineering, Erzincan, Turkey, ORCID: 0000-0002-7051-3368 eyyup.yildiz@erzincan.edu.tr

internet, the studies both accelerated and became easier to spread. As a result of these, artificial intelligence and image processing studies have increased rapidly and become popular. Thanks to the computers with high processing power used today, a wide variety of neural networks are used. Artificial intelligence has been showing its presence in almost every field very frequently in recent years. In fact, it is currently used in many industrial tools. This not only makes life easier, but also brings new fields of work. Artificial intelligence; It is divided into many sub-branches such as machine learning, deep learning, convolutional neural networks and divided into sections that are deepened in their own way.

Machine learning has started to take place in all areas of life. It has become widespread in many areas such as friend suggestions, movie suggestions, music suggestions in social networks, as well as phone applications and camera technologies. It has been a serious help in the development and dissemination of deep learning algorithms in areas such as translating speech into writing and extracting meanings from texts (LeCun, Bengio, & Hinton, 2015). With deep learning, operations such as image processing, object recognition and classification have gained a great momentum. It has become a frequently preferred method due to its high performance and the presence of subnets that deepen more than artificial neural networks (Wu, Liu, & Liu, 2019). Due to its good results in processing high-dimensional data and solving complex problems, its use has increased and is being used in many scientific fields and sectors (Farabet, Couprie, Najman, LeCun, & intelligence, 2012; Krizhevsky, Sutskever, & Hinton, 2012; Sutskever, Vinyals, & Le, 2014). Deep learning methods are used even in increasing the number of images and data augmentation processes (Mikołajczyk & Grochowski, 2018).

Convolutional neural networks are a deep learning architecture that tries to make it possible for living things to work similarly to their natural vision perceptions (Gu et al., 2018). CNN studies dating back to earlier times were developed and LeCun et.al. It was introduced in 1989 in his work on the classification of manuscripts (LeCun et al., 1989). It is frequently used because of the success of learning better by supporting deep learning structures with convolutional neural networks and the success of artificial intelligence in the classifier task in finding the best network path. It has become frequently preferred in subjects such as classification, object detection and recognition, especially thanks to its high-successful results on image data. In a classical CNN model, there is an input layer where the data is entered. Here, the data is provided to the network. Then it enters the convolutional layer and the features extracted from the images are scaled down by Max pooling. This is useful for a more in-depth feature extraction. Then the data going to Flatten and Fully Connected layers gives an output. This output gives the classification process. The general architecture of the CNN model is given in Figure 1.

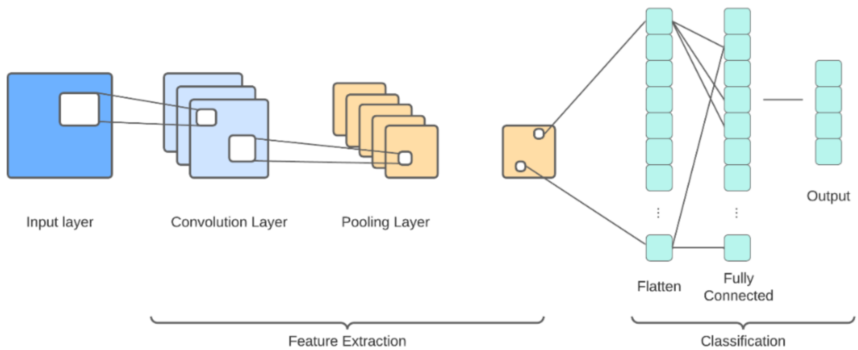


Figure 1: General structure of CNN architecture

In the study;

- Image classification is done. Here, it is aimed to classify images using a multi-class data set. The images are 150x150 in size and contain 25 thousand images. Test and Train parts were predetermined by the people who prepared the data set.

- The aim of the study is to conclude with the highest performance and the highest accuracy. In order to do this, many models have been tried and evaluated in an advisory context on how to choose the best method. Although there are different methods to find the best results, it is a recommendation made by the authors within the scope of the study. It is normal to have different approaches.

2. Related Works

In (Wan, Zeiler, Zhang, Le Cun, & Fergus, 2013), the cifar-10 dataset was studied using CNN. In the study, different models were tried and the parameters were optimized to get the best results. For this purpose, it has been analyzed by testing different optimization algorithms. Error rate was used as performance metrics and compared in this way. The best result was obtained with the Regularization of Neural Networks using DropConnect method with an Error rate of 0.22% (T. Guo, Dong, Li, & Gao, 2017).

Image processing and classification studies are also frequently performed on medical images. It is widely used with developing artificial intelligence algorithms and medical imaging instruments. In this study, a study was carried out on lung images. The CNN model was used to classify a lung disease. It is mentioned that a good result was obtained because the CNN model was designed by themselves. Precision and recall values were used as performance metrics. Precision and recall values have been completed with an average of approximately 90% success on the basis of classes (Li et al., 2014).

This study presents a comparative analysis using many data sets and many CNN models. The data sets used in the study are MNIST, CIFAR10, CIFAR100. The models used are GoogleNet, Resnet50 and AlexNet. Various classes are available in the datasets, such as household items, animals, tools. While performing these multiple classification processes, it was made within the framework of a wide comparison. The success rates of each class are given one by one and the prediction accuracy made by the CNN models is compared and the results are given (Sharma, Jain, & Mishra, 2018).

A study was presented using CNN models on hyperspectral images. In the study, a CNN model proposed by them was presented. Classification processing was performed with a model with 11 convolutional layers. Three different data sets were used to compare the method. It is said that it gives better results than previous CNN architectures. Working metrics were compared on accuracy and loss values and evaluated separately for each class. Comparisons were made in the form of CNN model and other methods. The results obtained are on the basis of class and it seems that there are high values (Yu, Jia, & Xu, 2017).

Al-Doski et. al., datasets created by remote sensing were studied. A classification was made on the images related to the land cover. It is aimed to evaluate the image sets used in the study in a process and to have high accuracy of the results to be obtained. K-means and SVM methods were used in the study (Al-Doski, Mansorl, & Shafri, 2013).

In (Rahimzadeh, Parvin, Safi, & Mohammadi, 2021), higher results were tried to be obtained by applying ResNet, DensNet and Inception models. Intel image classification data set was used as the data set. In the study, 224x224 resolution images and 512x512 images were handled in two different ways. It is said that the best scores are given. In the study, the results obtained over 80% in validation values are shared. and it is said that the accuracy rates in the models have been increased by around 2-8%.

Intel image classification data set was used in the study. A neural network called MinorNet has been proposed as a method. The data sets consist of multiple classes and the evaluation criteria to be made include the results of this multiple class. In addition, the study emphasizes its superiority over other methods. In the created CNN model, the validation success rate on high parameters was 88.3% (S. Guo, Ni, Xing, Liu, & Ni, 2021).

3. Materıal And Method

In the study, in general, a target was determined in terms of determining the basic materials by establishing small convolutional neural networks. Therefore, it is possible to see the best results, if necessary, optimizations are made by working on convolutional neural networks, which will be established in a small structure for each model. The general template of the study was made as given in Figure 2. In the pre-processing processes, a data generator was used and zoom_range=0.2, width_shift_range=0.2, shear_range=0.2, height_shift_range=0.2, horizontal_flip=True, vertical_flip=True, rotation_range=40 additions were made.

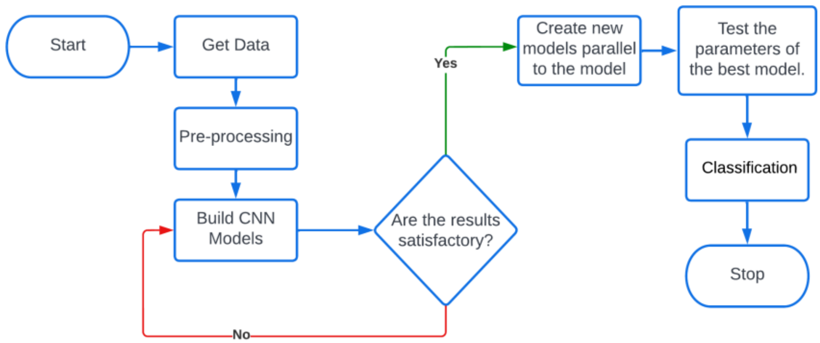


Figure 2: Flowchart of the study

After the data are collected, the best CNN model must be found in order to be able to classify. Therefore, it is necessary to compare the train and validation values obtained starting from the smallest models, and at some point, the better model should be selected and expanded, or a new model close to its equivalent should be added and compared. In order to do this, computers with high processing power must be used. In the study, Nvidia 1660Ti 6GB was used as GPU and Intel i7 9750H was used as CPU.

3.1 Dataset

The dataset used in the study is the Intel Image Classification dataset (Kaggle). There are 6 classes in total in the data set and these are; They are divided into buildings, forest glacier, mountain, sea, street. The data is available in zip file on the Kaggle site. There are folders reserved for 14 thousand trains, 3 thousand tests and 7 thousand predictions in the data set. The resolution of the images is determined as 150x150. The folder method has been chosen as the labeling format. The dataset was selected for classification using images in a similar category containing 6 classes. No visual changes were made to the data in the dataset. No noises have been added or subtracted and have been used purely.

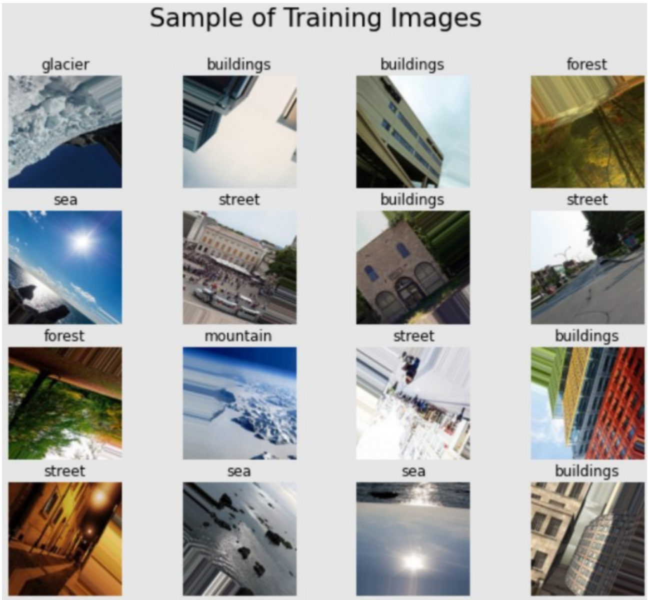


Figure 3: Some sample images in the dataset

3.2 Convolutional Neural Network (CNN)

There are many CNN architectures used in the studies. Basically, all CNN architectures consist of three basic layers. These; The convolutional layer is the pooling and fully connected layers. If modeled mathematically i. layer k. Their values at (i,j) position in the feature map are given in formula 1.

$$z_{i,j,k}^{l} = w_k^{l^T} x_{i,j}^{l} + b_k^{l} \tag{1}$$

$w_k^{l^T}$ and b_k^lused here are kth is the weight vector and bias term of the filter. The activation value of the $z_{i,j,k}^l$ convolutional property can be formulated as follows.

$$a_{i,j,k}^l = a(z_{i,j,k}^l) \tag{2}$$

In CNN architectures, a formation is shown within the framework of general layers. In this way, the data goes to the conclusion with an effort to extract a feature as it passes through each layer. The Lenet-5 architecture, which is one of the oldest architectural structures, and its features are given in Figure 4 (Gu et al., 2018).

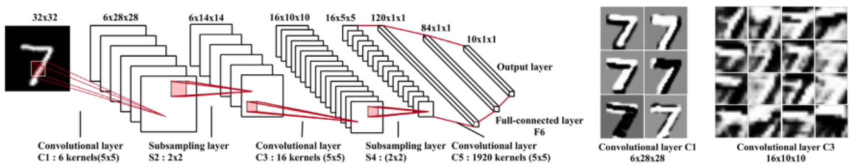


Figure 4: Lenet-5 Network and Learned Features

Basically, the basic features used in the CNN architecture were used in the study. Convolutional layer, Batch Normalization (Ioffe & Szegedy, 2015), Dropout (G. E. Hinton, Srivastava, Krizhevsky, Sutskever, & Salakhutdinov, 2012) and stride are used. In addition to these, many activations such as ReLU, ELU, SELU, Tanh, Softplus, Softmax (Bishop, 1995; Dugas, Bengio, Bélisle, Nadeau, & Garcia, 2000; G. E. Hinton et al., 2012; Lin & Shen, 2018; M. D. Zeiler et al., 2013) functions and optimizers such as; Adam, Adadelta, Adagrad, Adamax, Nadam, RMSprop, SGD (Bottou, 2010; Dozat, 2016; Duchi, Hazan, & Singer, 2011; G. Hinton, Srivastava, N., and Swersky, K., 2012; Kingma & Ba, 2014; M. D. J. a. p. a. Zeiler, 2012), are used. A wide variety of structures and parameters were used in the study, and so many structures were created. It is not possible to conduct such a large-scale architectural experiment for every dataset. Because in datasets with very large resolution and larger images, training of a structure can take tens of hours depending on the hardware used. It may not be possible to make such a wide range of architectural comparisons for every dataset. In order to make such comparisons, computers with either low resolution or very powerful hardware should be used. In addition, the architecture used in the study and to be concluded cannot be used as the best architecture. This is just an approximation and can be useful in building small scale architectures on low systems. Because the relatively short training periods and the corresponding performance results allow a usable structure.

Table 1: Stage-1 CNN Architectures

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Conv2D	16	24	32	48	64	128
Kernel_Size	3	3	3	3	3	3
Activation	ReLU**	ReLU**	ReLU**	ReLU**	ReLU**	ReLU**
Dropout	0.4	0.4	0.4	0.4	0.4	0.4
Conv2D	24	32	48	64	128	128
Kernel_Size	3	3	3	3	3	3
Activation	ReLU**	ReLU**	ReLU**	ReLU**	ReLU**	ReLU**
Dropout	0.4	0.4	0.4	0.4	0.4	0.4
Dense	128*	128*	128*	128*	128*	128*
Dropout	0.4	0.4	0.4	0.4	0.4	0.4
Dense	6	6	6	6	6	6
Activation	Softmax	Softmax	Softmax	Softmax	Softmax	Softmax
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam

* After this layer, BatchNormalization was applied.

** After this layer, BatchNormalization and Max Pooling were applied.

3.3 Applied CNN Models

The architectures used in the study were selected in terms of providing the best performance in the least amount of time. The table of these architectures is given below. The architectures used in the study were separated from each other as stages, and the model with the best results was expanded and new stages were created by making additions or changes. The architectures used for Stage-1 are given in Table 1. Each model was run at 20 epochs. As given in the starred note, there are layers with and without Batch Normalization and Max Pooling. First of all, experiments were started using small convolutional layers. The reason why the last drizzle layers are 6 is because a 6-class dataset is used. In addition, categorical crossentropy is used in all models as a loss function.

Table 2: Stage-2 CNN Architectures

	Model 7	Model 8	Model 9	Model 10	Model 11

Conv2D	64	64	64	64	64
Kernel_Size	3	3	3	3	3
Activation	ReLU**	ReLU**	ReLU**	ReLU**	ReLU**
Dropout	0.4	0.4	0.4	0.4	0.4
Conv2D	128	128	128	128	128
Kernel_Size	3	3	3	3	3
Activation	ReLU**	ReLU**	ReLU**	ReLU**	ReLU**
Dropout	0.4	0.4	0.4	0.4	0.4
Dense	16*	32*	64*	128*	256*
Dropout	0.4	0.4	0.4	0.4	0.4
Dense	6	6	6	6	6
Activation	Softmax	Softmax	Softmax	Softmax	Softmax
Optimizer	Adam	Adam	Adam	Adam	Adam

* After this layer, BatchNormalization was applied.

** After this layer, BatchNormalization and Max Pooling were applied.

The architectures used for Stage-2 are given in Table 2. Each model is trained for 20 epochs. As seen in the star note, Batch Normalization and Max Pooling have been done. Since it is a 6-class dataset, the last layer must be 6. Categorical crossentropy is used in all models as loss function. In this section, an investigation has been made on the effect of the changes in the Dense layer.

Table 3: Stage-3 CNN Architectures

	Model 12	**Model 13**	**Model 14**	**Model 15**	**Model 16**
Conv2D	64	64	64	64	64
Kernel_Size	3	3	3	3	3
Activation	ReLU**	ReLU**	ReLU**	ReLU**	ReLU**
Dropout	0.1	0.2	0.3	0.4	0.5

Conv2D	128	128	128	128	128
Kernel_Size	3	3	3	3	3
Activation	ReLU**	ReLU**	ReLU**	ReLU**	ReLU**
Dropout	0.1	0.2	0.3	0.4	0.5
Dense	64*	64*	64*	64*	64*
Dropout	0.1	0.2	0.3	0.4	0.5
Dense	6	6	6	6	6
Activation	Softmax	Softmax	Softmax	Softmax	Softmax
Optimizer	Adam	Adam	Adam	Adam	Adam

* After this layer, BatchNormalization was applied.

** After this layer, BatchNormalization and Max Pooling were applied.

The architectures used for Stage-3 are given in Table 3. Each model is trained for 20 epochs. As seen in the star note, Batch Normalization and Max Pooling have been done. Since it is a 6-class dataset, the last layer must be 6. Categorical crossentropy is used in all models as loss function. In this section, it is tried to make an examination on the effect of the changes when Dropout is made.

Table 4: Stage-4 CNN Architectures

	Model 17	**Model 18**	**Model 19**	**Model 20**	**Model 21**	**Model 22**
Conv2D	64	64	64	64	64	64
Kernel_Size	3	3	3	3	3	3
Activation	ReLU **	ReLU **	ReLU **	ReLU **	ReLU **	ReLU **

Dropout	0.1	0.1	0.1	0.1	0.1	0.1
Conv2D	128	128	128	128	128	128
Kernel_Size	3	3	3	3	3	3
Activation	ReLU **	ReLU **	ReLU **	ReLU **	ReLU **	ReLU **
Dropout	0.1	0.1	0.1	0.1	0.1	0.1
Dense	64*	64*	64*	64*	64*	64*
Dropout	0.1	0.1	0.1	0.1	0.1	0.1
Dense	6	6	6	6	6	6
Activation	Softmax	Softmax	Softmax	Softmax	Softmax	Softmax
Optimizer	SGD	RMSprop	Adadelta	Adagrad	Adamax	Nadam

* After this layer, BatchNormalization was applied.

** After this layer, BatchNormalization and Max Pooling were applied.

The architectures used for Stage-4 are given in Table 4. Each model is trained for 20 epochs. As seen in the star note, Batch Normalization and Max Pooling have been done. Since it is a 6-class dataset, the last layer must be 6. Categorical crossentropy is used in all models as loss function. In this section, an investigation has been made on the effect of the changes made with the optimizer.

Table 5: Stage-5 CNN Architectures

	Model 23	**Model 24**	**Model 25**	**Model 26**	**Model 27**	**Model 28**	**Model 29**
Conv2D	64	64	64	64	64	64	64
Kernel_Size	3	3	3	3	3	3	3
Activation	Softplus**	Softmax **	ReLU **	Softsign **	Tanh **	Selu **	Elu**
Dropout	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Conv2D	128	128	128	128	128	128	128
Kernel_ Size	3	3	3	3	3	3	3
Activation	Softplus**	Softmax **	ReLU **	Softsign **	Tanh **	Selu **	Elu**
Dropout	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Dense	64*	64*	64*	64*	64*	64*	64*
Dropout	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Dense	6	6	6	6	6	6	6
Activation	Softmax	Softmax	Softmax	Softmax	Softmax	Softmax	Softmax
Optimizer	Adadelta	Adadelta	Adadelta	Adadelta	Adadelta	Adadelta	Adadelta

* After this layer, BatchNormalization was applied.

** After this layer, BatchNormalization and Max Pooling were applied.

The architectures used for Stage-5 are given in Table 5. Each model is trained for 20 epochs. As seen in the star note, Batch Normalization and Max Pooling have been done. Since it is a 6-class dataset, the last layer must be 6. Categorical crossentropy is used in all models as loss function. In this section, an attempt is made to examine the effect of changes made with activation.

Testing and validation results of all models are given and discussed in Chapter 4. Detailed results are given in this section. In addition, it was stated why the models were chosen and which model was superior. The criteria for the superiority of the models were made by considering both the formations in the graphics and the duration of the train.

Layer (type)	**Output Shape**	**Param #**
conv2d_1 (Conv2D)	(None, 148, 148, 64)	1792
batch_normalization_1 (Batch)	(None, 148, 148, 64)	256
max_pooling2d_1(MaxPooling2)	(None, 74, 74, 64)	0
dropout_1 (Dropout)	(None, 74, 74, 64)	0
conv2d_2 (Conv2D)	(None, 36, 36, 128)	73856
batch_normalization_2 (Batch)	(None, 36, 36, 128)	512
max_pooling2d_2(MaxPooling2)	(None, 18, 18, 128)	0
dropout_2 (Dropout)	(None, 18, 18, 128)	0

conv2d_3 (Conv2D)	(None, 8, 8, 256)	295168
max_pooling2d_3 (MaxPooling2)	(None, 8, 8, 256)	0
flatten_1 (Flatten)	(None, 4096)	0
dense_1 (Dense)	(None, 64)	262208
batch_normalization_3 (Batch)	(None, 64)	256
dropout_3 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 6)	390

Total params: 634,438
Trainable params: 633,926
Non-trainable params: 512

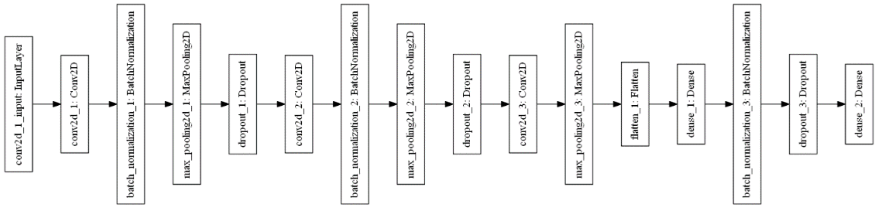


Figure 5: Last CNN Model

4. Experiments and Discussion

The results of the created models are explained in this section. The basic features of the models are shown and discussed on the figures. In order to measure their performance, train loss and validation loss values are shown and evaluated as well as train and validation values.

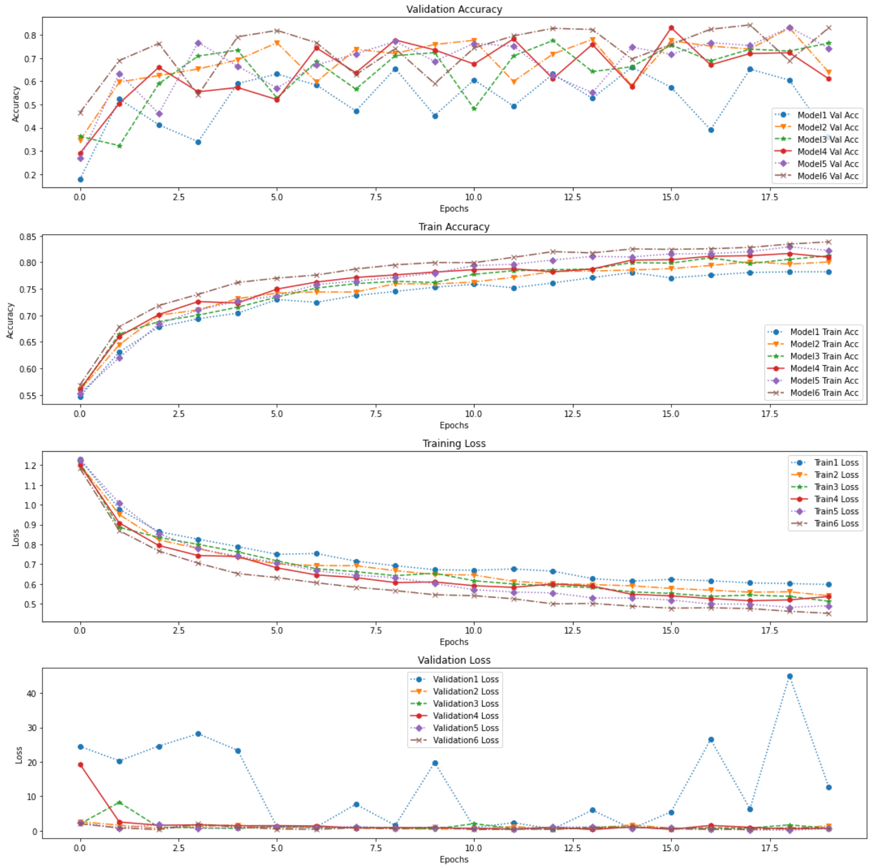


Figure 6: Train, Train loss, Validation and Validation Loss Values of Models 1 to 6.

As can be seen in Figure 6, many models have been compared. It has been clearly seen that Model 6 has higher success in both test and validation values. In addition, when we look at the loss values, Model 6 has achieved success at lower loss values. Model 1 appears to have had a very unbalanced training. Looking at the models here, it is seen that Model 5 is also in a good level. However, since it was not thought to provide enough performance, Model 6 was continued by changing and expanding. One of the important things here is the balanced ups and downs in the graphics. The fact that these fluctuations are low, that is, stable, can also give a clue about the accuracy of the model. For this reason, it should be taken into account when considering the training of the models and the fluctuations. On the Model 6, each epoch takes an average of 85 seconds.

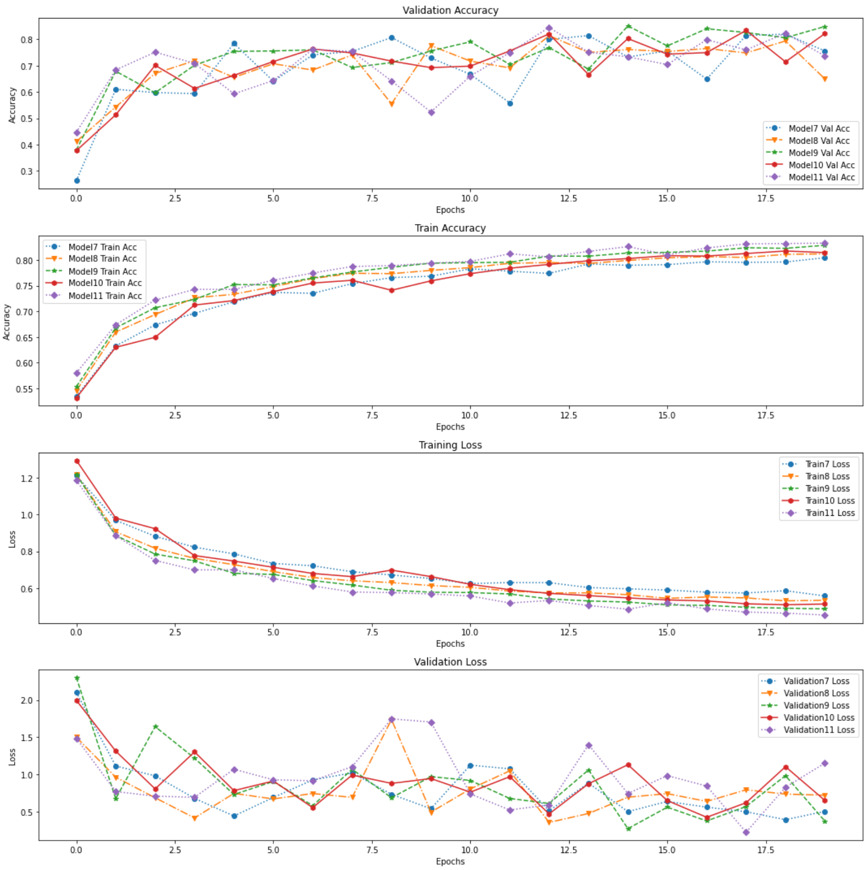


Figure 7: Train, Train loss, Validation and Validation Loss Values of Models 7 to 11.

As seen in Figure 7, train and train loss values have a less fluctuating structure. As a result of the training, it is seen that the validation values are more fluctuating. This means that it would be more accurate to choose the best validation accuracy values, which generally give us better results when choosing. For this reason, all differences are taken into account when choosing the best model. Model 11's loss values seem to be good. However, it is seen that the validation loss and validation accuracy values of Model 9 are better. This creates the opinion that the selection is stronger than Model 9 and that the selection will be more accurate. There are two models close to each other that can be selected here. Going both ways would not be a bad choice. However, since validation accuracy is considered more important, it was continued by paying attention to this when choosing.

In the training results given in Figure 8, it is possible to make a selection in a similar way. Validation loss values approaching zero are seen for each model. It is seen that the train and train loss values are smooth without waves and are now aligned. This indicates a development in the direction that progress is correct. At the same time, it is seen that the success rates have already exceeded 80%. While training, it was run at relatively low epochs instead of high epochs, and for this reason, it was possible to try many models. Looking at this figure, without much hesitation, the Model 12 is clearly better. For this reason, this model has been developed. It is seen that the training process is correct and there is slight fluctuation in the validation processes. However, this may also be related to the unstable situation in the operating performance of the computer. Therefore, the opinion that the model of every fluctuation is wrong may not be correct. It is quite normal to have such situations when choosing a model. In addition, these fluctuations are likely to occur even in computers with high processing power. Because it cannot be expected to see a work at the same frequency values at every moment. Considering this figure, model 12 is clearly ahead in terms of train and train loss in model selection. It is also clearly seen that it is ahead for validation accuracy. Even if the validation accuracy is low with a very small margin, it would be more accurate to choose this model. Because it is seen to be good in many aspects and it will be a predictable prediction that it will give better results at large epoch values.

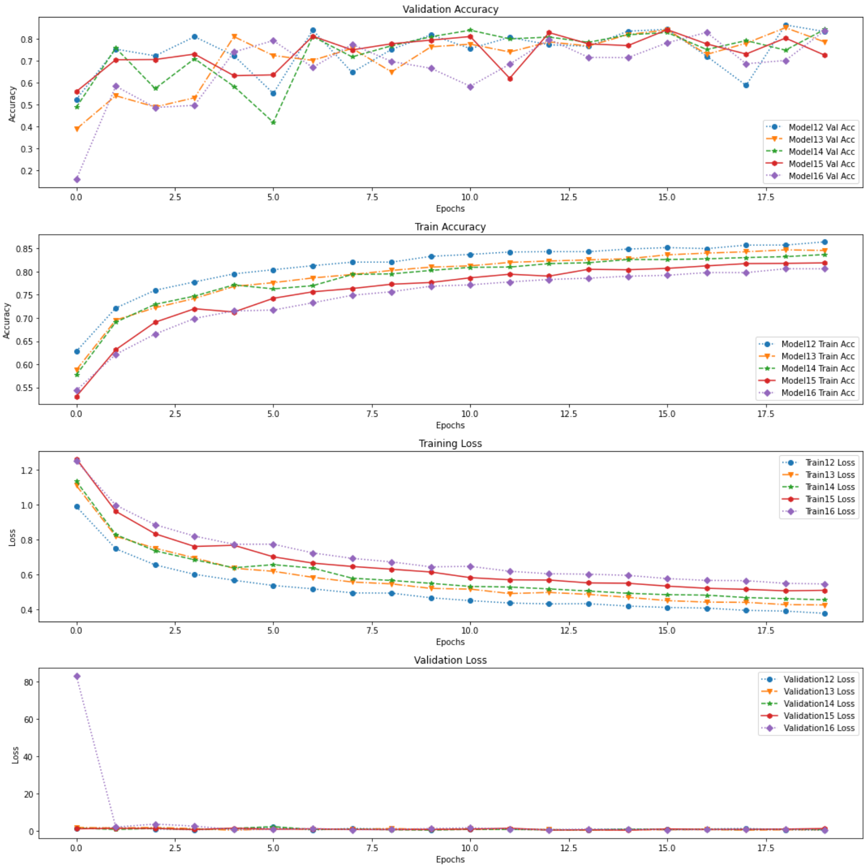


Figure 8: Train, Train loss, Validation and Validation Loss Values of Models 12 to 16.

Similarly, in Figure 9, it is aimed to choose a model with low volatility, low loss values and high validation accuracy. Model 17 is the first model to be eliminated as it is obviously a model with too much volatility and low results. When all values are considered, it is seen that Model 19 is the best. Because train values and train loss values are obviously the best. It is the model that sees the highest value in validation accuracy.

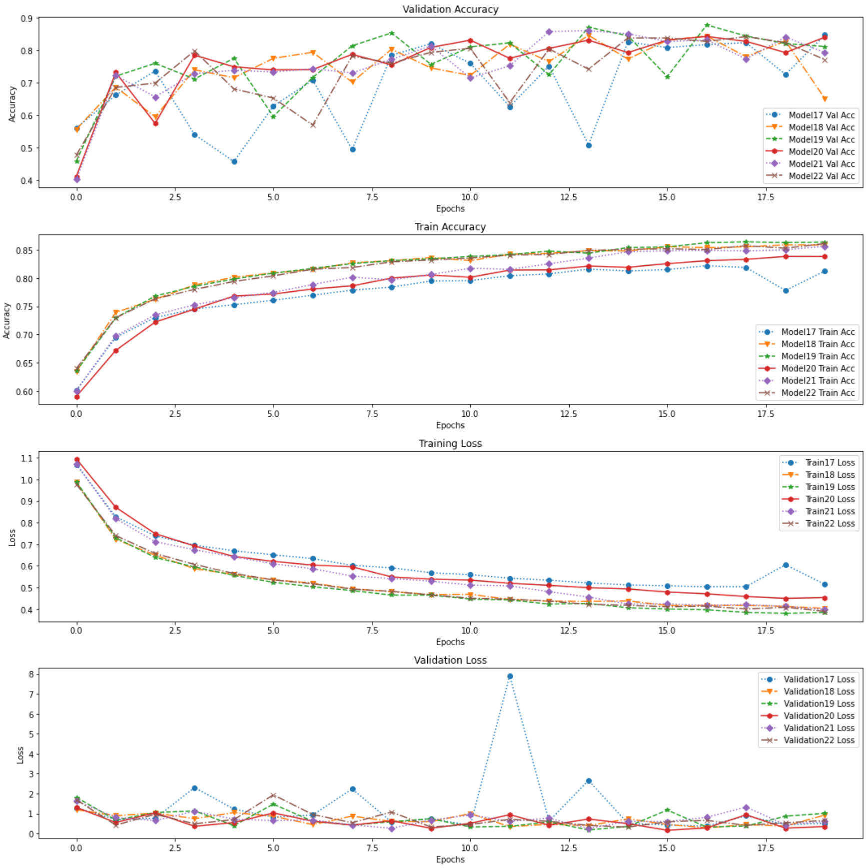


Figure 9: Train, Train loss, Validation and Validation Loss Values of Models 17 to 22

The final model to be compared is given in Figure 10. The thing to note here is that although it is thought to be seen as very wavy, the places with very small value differences appear big because they are shown on a large figure. For example, although the loss values in the validation loss section are between 3 and 0.5, it seems that there are big fluctuations because it is in a large display. While naturally preferred, it is essential to choose the models that give the least fluctuation and the best results. In many respects, the Model 25 seems to give the best results.

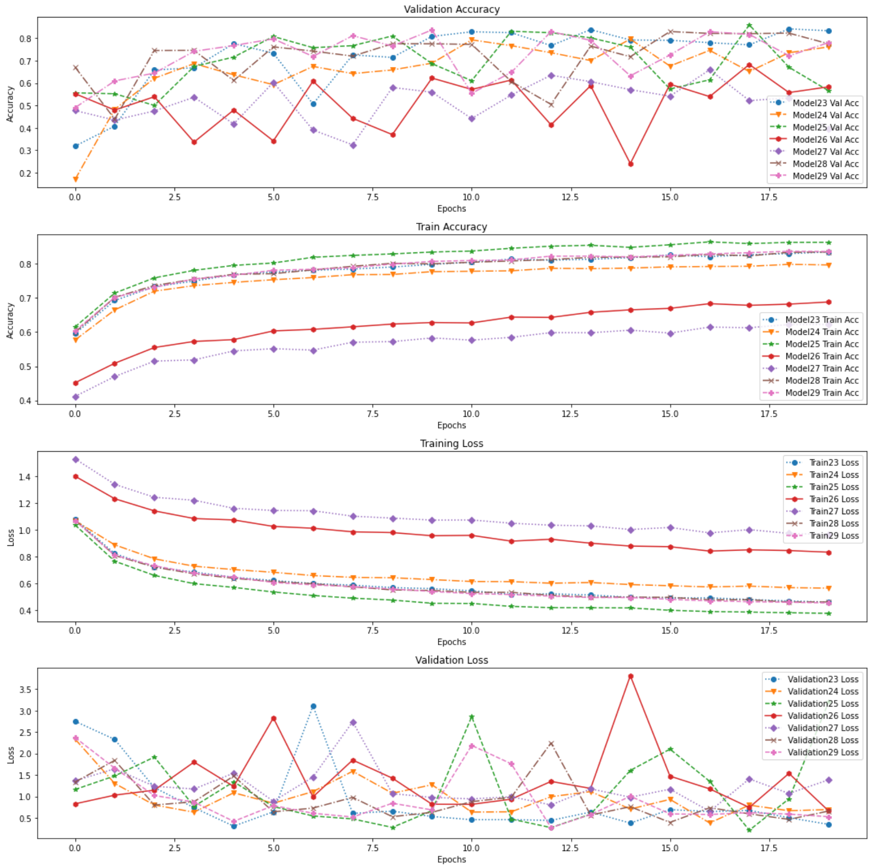


Figure 10: Train, Train loss, Validation and Validation Loss Values of Models 23 to 29

Finally, Figure 11 shows the results. Here, by considering Model 25, an architecture has now been created using an expanded convolutional neural network. The best results were also performed on this model. Train accuracy was approximately 91%, while validation accuracy reached 89.23%. The best results obtained were also carried out on this model. Here, the layers in the architecture have been expanded and the number of epochs has been increased to 50. It is seen that the studies carried out are systematic and accurate. Loss values were reduced below 0.5%. This proves that the model works well.

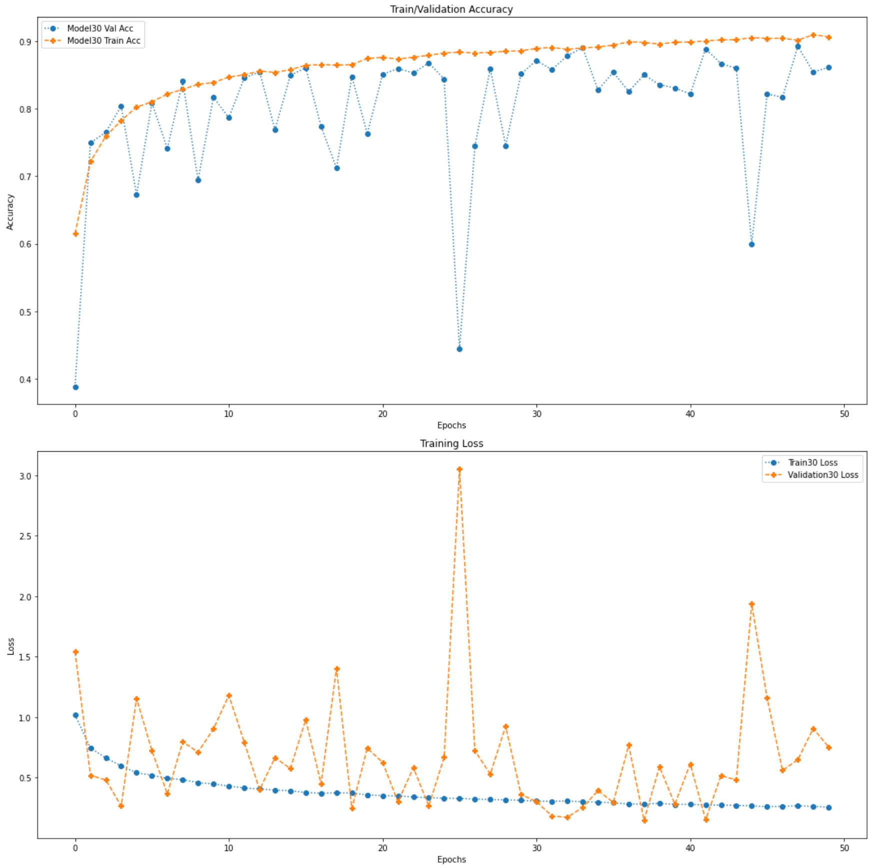


Figure 11: Last and Best Model

5. Conclusion

In the study, the results related to how a CNN model should be applied for a classification process were investigated. By using the smallest convolutional neural networks for the establishment of the models, it is provided to expand by adding extra new layers. With every good result found, an analysis was made to find better models by switching to a new stage, adding layers and changing some parameters. These processes have been made entirely by trial and error and have been expanded in a systematic structure. The results obtained and the architectures to be installed are likely to vary depending on the dataset. However, similar to this systematic structure, it can be applied with any dataset. It is normal to see changes with the computer used in the study and the models created. In addition, it should be said that a subjective selection is made while examining the models. It would be more accurate to have enough training time to see each model yield fully. However, when a simple and result-oriented architecture should be used, it would be beneficial to use such a method. Because it is seen that the results obtained are both trained in short periods of time and classification is made at high achievements. In order to get a better efficiency from this result, a more detailed architectural examination is required. In this way, the performance values found in the study can be improved and better results can be achieved.

In the study, it is aimed to obtain a result in the way of comparison and observation by considering many models both in terms of architecture and parameters. It can be said that the results are satisfactory enough. Train accuracy was 91% and validation accuracy was 89.23%, and a 6-class data set was classified. By expanding the models further, training can be done at higher times and better results can be obtained. However, the focus of this study is to establish a fast architecture and obtain high-performance results with low training times. The main purpose of the study was made within the framework of this focus. In addition, it is seen that the results obtained prove this.

6. References

Al-Doski, J., Mansorl, S. B., & Shafri, H. Z. M. J. D. o. C. E., Faculty of Engineering, University Putra, Malaysia. (2013). Image classification in remote sensing. *3*(10).

Arf, C. J. A. Ü.-Ö. Y. H. K. (1959). Makineler Düşünebilir mi ve Nasıl Düşünebilir? (1), 91-103.

Bishop, C. M. (1995). *Neural networks for pattern recognition*: Oxford university press.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010* (pp. 177-186): Springer.

Dozat, T. (2016). Incorporating nesterov momentum into adam.

Duchi, J., Hazan, E., & Singer, Y. J. J. o. m. l. r. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *12*(7).

Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., & Garcia, R. J. A. i. n. i. p. s. (2000). Incorporating second-order functional knowledge for better option pricing. *13*.

Farabet, C., Couprie, C., Najman, L., LeCun, Y. J. I. t. o. p. a., & intelligence, m. (2012). Learning hierarchical features for scene labeling. *35*(8), 1915-1929.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., . . . Cai, J. J. P. r. (2018). Recent advances in convolutional neural networks. *77*, 354-377.

Guo, S., Ni, Y., Xing, K., Liu, Y., & Ni, W. (2021). *MinorNet: A Lightweight neural network for battlefield scene* *classification.* Paper presented at the 2021 14th International Symposium on Computational Intelligence and Design (ISCID).

Guo, T., Dong, J., Li, H., & Gao, Y. (2017). *Simple convolutional neural network on image classification.* Paper presented at the 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA).

Hinton, G., Srivastava, N., and Swersky, K. (2012). Lecture 6d - a separate, adaptive learning rate for each connection. *Slides of Lecture Neural Networks for Machine Learning,.*

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. J. a. p. a. (2012). Improving neural networks by preventing co-adaptation of feature detectors.

Ioffe, S., & Szegedy, C. (2015). *Batch normalization: Accelerating deep network training by reducing internal covariate shift.* Paper presented at the International conference on machine learning.

Kaggle. Intel Image Classification https://www.kaggle.com/datasets/puneet6060/intel-image-classification (Last Access: 18.08.2022)

Kingma, D. P., & Ba, J. J. a. p. a. (2014). Adam: A method for stochastic optimization.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Proc. Advances in Neural Information Processing Systems 25. In.

LeCun, Y., Bengio, Y., & Hinton, G. J. n. (2015). Deep learning. *521*(7553), 436-444.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. J. A. i. n. i. p. s. (1989). Handwritten digit recognition with a back-propagation network. *2*.

Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D. D., & Chen, M. (2014). *Medical image classification with convolutional neural network.* Paper presented at the 2014 13th international conference on control automation robotics & vision (ICARCV).

Lin, G., & Shen, W. J. P. c. s. (2018). Research on convolutional neural network based on improved Relu piecewise activation function. *131*, 977-984.

Mikołajczyk, A., & Grochowski, M. (2018). *Data augmentation for improving deep learning in image classification problem.* Paper presented at the 2018 international interdisciplinary PhD workshop (IIPhDW).

Rahimzadeh, M., Parvin, S., Safi, E., & Mohammadi, M. R. J. a. p. a. (2021). Wise-srnet: A novel architecture for enhancing image classification by learning spatial resolution of feature maps.

Sharma, N., Jain, V., & Mishra, A. J. P. c. s. (2018). An analysis of convolutional neural networks for image classification. *132*, 377-384.

Sutskever, I., Vinyals, O., & Le, Q. (2014). Proc. Advances in Neural Information Processing Systems 27. In.

Turing, A. M., & Haugeland, J. J. T. T. T. V. B. a. t. H. o. I. (1950). Computing machinery and intelligence. 29-56.

Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., & Fergus, R. (2013). *Regularization of neural networks using dropconnect.* Paper presented at the International conference on machine learning.

Wu, H., Liu, Q., & Liu, X. J. C. M. C. (2019). A review on deep learning approaches to image classification and object segmentation. *60*, 575-597.

Yu, S., Jia, S., & Xu, C. J. N. (2017). Convolutional neural networks for hyperspectral image classification. *219*, 88-98.

Zeiler, M. D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q. V., . . . Dean, J. (2013). *On rectified linear units for speech processing.* Paper presented at the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing.

Zeiler, M. D. J. a. p. a. (2012). Adadelta: an adaptive learning rate method.

CHAPTER III

Bibliometric Analysis of Ant Colony Optimization Algorithm (1992-2024)

Ahmet AKKAYA[1]
Cemil KÖZKURT[2]

1. Introduction

Optimization is a technique that seeks solutions for a system to perform optimally under certain constraints (Vanderplaats, 1988) . Traditional optimization methods are inadequate for nonlinear, dynamic and large-scale problems (Blum & Roli, 2003). In these cases, metaheuristic optimization techniques provide an alternative approach by providing fast and efficient approximate solutions (Eid et al., 2022) . The most common metaheuristic algorithms include Simulated Annealing (SA), Genetic Algorithm (GA), Grey Wolf

[1] *Bandırma Onyedi Eylül University, Bandırma, Turkey, aakkaya@bandirma.edu.tr, ORCID: 0000-0003-4836-2310*
[2] *Bandırma Onyedi Eylül University, Bandırma, Turkey, ckozkurt@bandirma.edu.tr, ORCID: 0000-0003-1407-9867*

Optimization Algorithm (GWO), Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) (Akkaya et al., 2023).

ACO is an algorithm inspired by the foraging behavior of natural ant colonies (Dorigo & Stützle, 2004). The basic logic of ACO is that by using pheromone trails, ants can collectively find optimal paths (Durgut, 2021) . This mechanism produces effective solutions, especially in combinatorial and dynamic optimization problems.

The mathematical foundations of ACO are expressed by two main equations: pheromone update and probabilistic path selection. Equation 1 gives the pheromone update equation.

$$\tau_{ij} = (1 - \rho) \cdot \tau_{ij} + \Delta\tau_{ij} \qquad \text{Equation (1)}$$

It's here:

- τ_{ij} is the amount of pheromone on path (i, j).
- ρ is the rate of pheromone evaporation and ensures that old information is erased.
- $\Delta\tau_{ij}$ represents the amount of new pheromone released by the ants.

The pheromone update ensures that the trails left by the ants evaporate over time and this plays an important role in the effort to reach the optimal solution (Dorigo & Gambardella, 1997). Equation 2 gives the probabilistic path selection equation.

$$P_{ij} = \frac{(\tau_{ij})^{\alpha} \cdot (\eta_{ij})^{\beta}}{\sum_{k \in allowed_j} (\tau_{ik})^{\alpha} \cdot (\eta_{ik})^{\beta}} \qquad \text{Equation (2)}$$

- η_{ij} represents the heuristic value of the path.
- α and β are parameters that determine the importance of the pheromone effect and heuristic information, respectively.

The possibility of path selection allows ants to discover better paths when searching for solutions. Equation 1 and Equation 2 allow ACO to find better solutions by avoiding local optima (Socha & Dorigo, 2008).

As with any meta-heuristic algorithm, ACO has a number of advantages and disadvantages. The most important advantages of ACO are flexibility, parallel search and adaptive. Flexibility, ACO is suitable for a wide range of optimization problems and provides effective solutions for combinatorial, continuous and multi-objective problems (Dorigo & Stützle, 2004). Parallel search, the colony structure enables a parallel search in the solution space, which increases the speed of finding solutions (Blum & Roli, 2003). Adaptive, its ability to adapt to changing conditions improves its performance in dynamic environments (Socha & Dorigo, 2008). The disadvantages are computational cost, local optimum and parameter tuning. Computational Cost, complex and large-scale problems may require high computational power and long processing times (Mavrovouniotis & Yang, 2015). Local Optimum, sometimes it can get stuck at the local optimum, which can make it difficult to reach the global optimum (Blum & Roli, 2003). Parameterization, it is important to set the parameters of the algorithm correctly, otherwise the performance of the algorithm is negatively affected (López-Ibáñez & Stützle, 2010).

ACO has been widely used in logistics and route optimization (Zhang et al., 2022) , data mining (Ganji & Abadeh, 2010) and feature selection (Ahmad et al., 2017) , multi-objective and continuous optimization (Emami Skardi et al, 2015) , network and communication systems (Ma et al., 2022) and artificial intelligence and robotics (Sandini et al., 2024) . In logistics and route optimization, it is used in the traveling salesman problem (Chaudhari & Thakkar, 2019), which is the problem of finding the shortest path between a set of cities, in the vehicle routing problem (Revanna & Al-Nakash, 2023)), which is the optimization of distribution networks and transportation systems. It is used in engineering designs (Kaveh & Talatahari, 2010) and (Mardoude et al., 2024) to improve the efficiency of energy systems. Internet data routing (Safavat & Rawat, 2021) and load balancing in communication networks (Wang et al., 2016) It is used in Artificial Intelligence and Robotics for pathfinding algorithms (Ming et al., 2021) of autonomous vehicles and trajectory (Madridano et al., 2021) planning of multi-robot systems.

In this study, the research on the Ant Colony Optimization (ACO) algorithm since its inception in 1992 is examined through bibliometric analysis. The study examines the development and prevalence of the algorithm over time by presenting the trends in the use of ACO with quantitative data. The next section provides information about the purpose of the study, the methods used and the data analyzed.

A. Methodology of the study

Bibliometric analysis is a quantitative method used to understand the dynamics of academic literature and assess research trends. This method was developed to systematically examine the density of publications, collaboration networks among authors, and citation relationships in a given scientific field (Donthu et al., 2021) . (Price, 2019) The "law of scientific growth" study laid the foundation for bibliometric analyses by demonstrating that scientific production increases geometrically over time. Later (Small, 1973) , developed citation analysis and revolutionized the visualization of conceptual clusters and knowledge diffusion in the literature. Today, bibliometric studies are carried out using software such as VOSviewer, CiteSpace to analyze large-scale data from databases such as Scopus and Web of Science (Aria & Cuccurullo, 2017) . These analyses provide important contributions in terms of evaluating academic performance, mapping research networks and examining the flow of information across disciplines.

B. Purpose of the research

The aim of this research is to examine a total of 20,740 studies on ACO in the Web of Science (WoS) database in the last 33 years (1992-2024) using the keywords "ant colony optimization algorithm" OR "ant colony algorithm" OR "ant colony system" OR "ant colony" by using bibliometric analysis method. This analysis aims to reveal the trends of research in the field of ACO, author collaborations, most cited studies and journal impacts. It also aims to draw the attention of researchers by identifying the development of ACO over time and the most remarkable aspects in the scientific field. The quantitative data obtained provide a valuable basis for better understanding the evolution of ACO research, key topics and important contributions to the field.

C. Data and analysis

Software such as CiteSpace, Bibliometrix, Gephi, HistCite, EndNote and VOSviewer are frequently used in bibliometric analysis. In this study, Web of Science (WoS) database was preferred to ensure reliability and comprehensive data analysis. The dataset of 20,740 publications obtained from WoS was analyzed using VOSviewer software, which stands out especially for visualization of scientific literature and network analysis. VOSviewer is an effective tool for literature and network analysis thanks to its user-friendly interface, free access and capacity to process large data sets. This software has revealed trends and research gaps in the field of optimization by mapping interactions, connections and network structures between scientific publications. In the study, the analysis using the keywords "ant colony optimization algorithm" OR "ant colony algorithm" OR "ant colony system" OR "ant colony" provided a comprehensive map of the scientific literature on ACO algorithms as of November 2024 and shed light on current trends in the field.

2. Findings

In this section, detailed analyses of different categories in the Web of Science (WoS) database are presented. Comprehensive data such as the distribution of publications over the years, document types, WoS index coverage, research profiles and the distribution of publication titles according to research categories are analyzed. In addition, the distribution of publications by country and publisher, and the diversity and distribution of research areas are evaluated. The meso- and micro-level distribution of citation topics and the classification of publications according to their languages are discussed. In the co-authorship analyses, collaborations at the researcher, institution and country levels were detailed, and the citation relationships of sources, authors and institutions were analyzed. The findings provide important data to understand the evolution of research in the ACO field, its place in the global scientific network and its dynamics. With these analyses, it is aimed to reveal more clearly the development and contributions of the field within the scientific community.

WoS categories distribution

As a result of the analysis, Figure 1 the distribution of publications according to WoS categories.

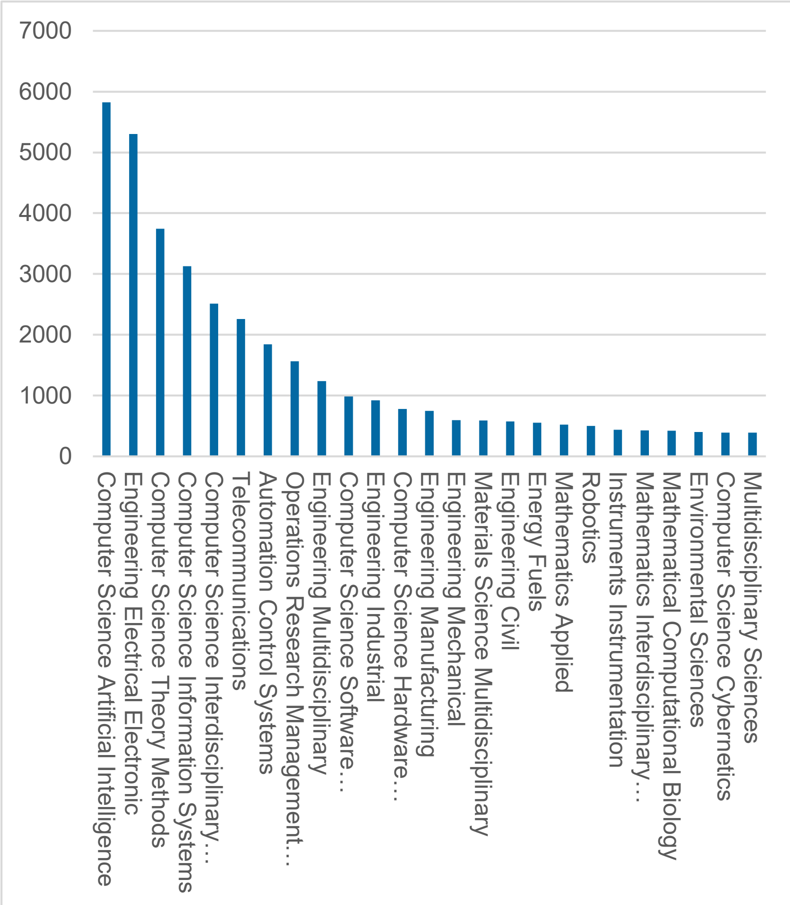


Figure1 : Distribution of publications according to WoS categories

When Figure 1 is analyzed, it is seen that computer sciences such as artificial intelligence, theory and methods, information systems, software engineering come to the forefront. Artificial Intelligence (5826), as the category with the highest number of records, shows that ACO algorithms have a wide range of applications in artificial intelligence applications. Theory and

Methods, with (3743) publications, shows that the theoretical foundations and methodological developments of ACO algorithms are intensively discussed in this field. Information Systems 3129 shows that these algorithms play an important role in information processing, data analysis and optimization processes. In terms of engineering applications, Electrical and Electronics Engineering (5302) publications and the density of optimization problems in this field explain the widespread use of ACO algorithms. Multidisciplinary Engineering, with (1236) publications, shows that ACO algorithms are an effective tool for solving problems that combine different engineering disciplines. Industrial and Mechanical Engineering, with (922 and 597) publications respectively, make significant contributions in areas such as manufacturing, process optimization and design. Operational Research and Management Science (1566), decision-making processes, resource allocation and optimization problems are concentrated in this category. Telecommunications (2257) publications with the effectiveness of ACO algorithm in network optimization, data transmission and design of communication systems. Robotics (502) with the effectiveness of these algorithms in robot motion planning, control and automation processes. Mathematical and Computational Biology (420) record with ACO algorithm is widely used in modeling and simulation of biological systems.

Distribution of publications according to years of publication

According to the findings obtained as a result of the analysis process, the distribution of publications according to years is given in Figure 2.

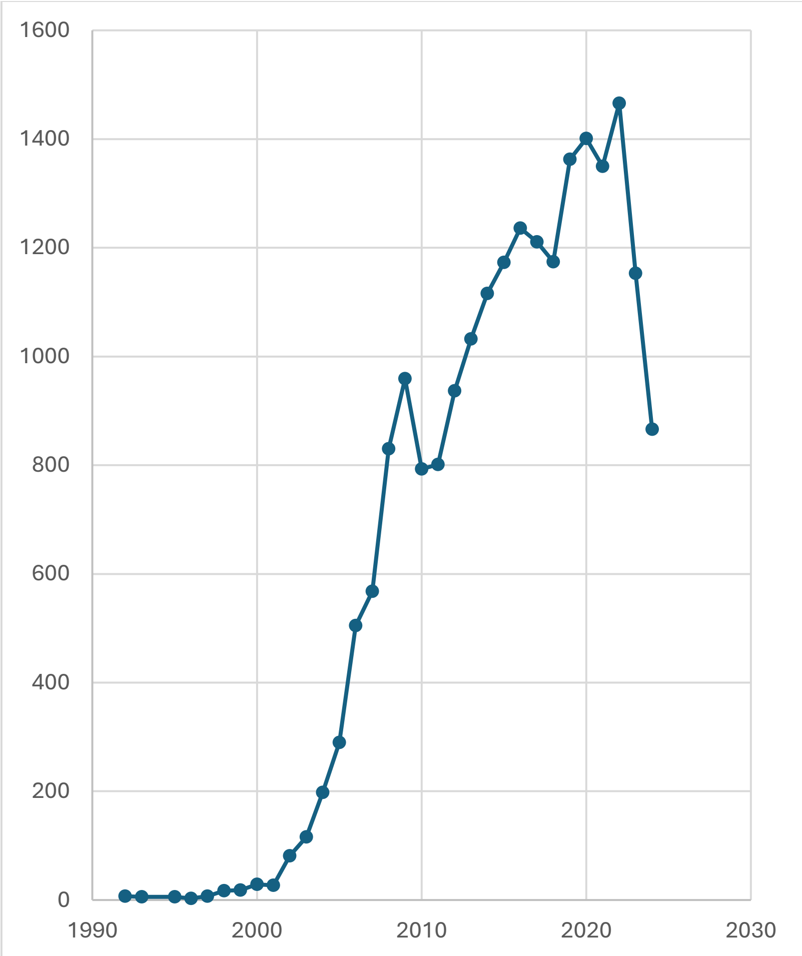


Figure2 : Distribution of publications by years

When Figure 2 is analyzed, it is seen that the number of publications towards the end of the 1990s (1992-1999) was quite low. In the years before the late 1990s, very few publications were made. It is observed that only a few publications were made in 1996, 1997 and 1993. This shows that research in the field of ACO was much more limited in this period due to its newness. As of the 2000s, an increase begins. Although it is seen that it started with 29 publications in 2000 and 27 publications in 2001, it still remains at low levels. A more significant growth is observed in 2005-2010. While there were 290 publications in 2005, this number increased to 793 by 2010. During this period, scientific research on ACOs started to become more widespread. In the 2010s, this increase continued to accelerate. It is seen that it gained momentum with 937 publications in 2012, 1032 in 2013, 1116 in 2014, and 1173 in 2015. In 2016 and 2017, the annual number of publications increased even more, reaching 1200. After 2017, a significant increase in the number of publications was observed. In 2020, it reached its peak with 1401 publications, followed by 1350 in 2021, 1466 in 2022 and 1153 in 2023. In this period, there was a significant increase in scientific production and research intensity, reaching 866 publications in 2024.

Distribution of publications by document type

Table 1 shows the distribution of 20,740 publications in the WoS database within the scope of ACO according to document types and the percentage of each type in total publications.

Table1 : Distribution of publications by document type

Document Types	Record Count	% of 20.740
Article	12243	59.031
Book	6	0.029
Book Chapters	311	1.5
Book Review	1	0.005
Correction	20	0.096
Early Access	167	0.805

Editorial Material	34	0.164
Letter	3	0.014
Meeting Abstract	20	0.096
News Item	1	0.005
Proceeding Paper	8384	40.424
Publication With Expression Of Concern	1	0.005
Retracted Publication	104	0.501
Retraction	14	0.068
Review Article	449	2.165
Software Review	1	0.005

Article 12,243 records, accounting for 59,031% of the total. This is the most common document type and shows that the majority of scientific research is published in article format. Conference Proceeding Paper (Proceeding Paper) 8,384 publications, accounting for 40,424% of the total. Conference proceedings also occupy an important place. This indicates that most of the research was presented at conferences and then published. Review Article 449 publications, Book 311 publications, Early Access 167 publications, Retracted Publication 104 publications, Editorial Material 34 publications, Correction 20 publications, Meeting Abstract 20 publications, Retraction 14 publications, Book 6 publications, Letter 3 publications, News Item 1 publication, Publication with Expression of Concern 1 publication, Software Review 1 publication.

Meso-distribution of citation topics of publications

As a result of the analysis, meso (intermediate level) distribution of publications according to citation topics is presented in Table 2 in order to reveal which fields of study come to the fore and which topics attract more attention.

Table 2 : Meso (medium level) distribution of publications according to citation topics

Citation Topics Meso	Record Count	% of 20.740
4.84 Supply Chain & Logistics	6894	33.24
4.13 Telecommunications	1936	9.335
4.61 Artificial Intelligence & Machine Learning	1131	5.453
4.29 Automation & Control Systems	1032	4.976
4.46 Distributed & Real Time Computing	998	4.812
4.18 Power Systems & Electric Vehicles	969	4.672

4.224 Design & Manufacturing	485	2.338
4.17 Computer Vision & Graphics	453	2.184
4.183 Transportation	419	2.02
4.47 Software Engineering	417	2.011
4.48 Knowledge Engineering & Representation	397	1.914
3.32 Entomology	373	1.798
6.153 Climate Change	271	1.307
4.237 Safety & Maintenance	211	1.017
4.116 Robotics	182	0.878
4.58 Wireless Technology	142	0.685
2.123 Protein Structure, Folding & Modeling	135	0.651
4.101 Security, Encryption & Encoding	113	0.545
4.206 Models Of Computation	103	0.497
6.11 Education & Educational Research	103	0.497
4.169 Remote Sensing	97	0.468
7.227 Manufacturing	97	0.468
7.63 Mechanics	96	0.463
8.212 Sensors & Tomography	93	0.448
2.244 Chemometrics	91	0.439

Table 2 shows that Supply Chain & Logistics is the most highly cited topic with 6,894 publications. Since supply chain and logistics research are key components of global trade and industrial production, it is a very popular and intensively researched area. Telecommunications ranks second with 1,936 publications. Telecommunications is a rapidly developing sector and continues to be an important research topic with digitalization and the proliferation of connected devices. Artificial Intelligence & Machine Learning ranks third with 1,131 publications. Artificial intelligence and machine learning is a rapidly growing field in the world of technology and science, finding applications in every sector. Automation & Control Systems 1.032 publications, Distributed &

Real Time Computing 998 publications, Power Systems & Electric Vehicles 969 publications, Design & Manufacturing 485 publications, Computer Vision & Graphics 453 publications, Transportation 419 publications, Software Engineering 417 publications, Knowledge Engineering & Representation 397 publications, Entomology 373 publications, Climate Change 271 publications, Safety & Maintenance 211 publications, Robotics 182 publications, Wireless Technology 142 publications, Protein Structure, Folding & Modeling 135 publications, Security, Encryption & Encoding 113 publications, Models of Computation 103 publications, Education & Educational Research 103 publications, Remote Sensing 97 publications, Manufacturing 97 publications, Mechanics 96 publications, Sensors & Tomography 93 publications, Chemometrics 91 publications.

Micro-distribution of citation topics of publications

Table 3 shows the classification of ACO publications according to micro-level, i.e. narrow-scale citation topics and the number of publications on each topic.

Table3 : Classification of publications according to micro-level citation topics

Citation Topics Micro	**Record Count**	**% of 20.740**
4.84.169 Particle Swarm Optimization	3818	18.409
4.84.471 Vehicle Routing Problem	1435	6.919
4.13.43 Wireless Sensor Networks	1167	5.627
4.84.401 Scheduling	1093	5.27
4.29.435 Multi Agent Systems	788	3.799
4.46.85 Cloud Computing	635	3.062
4.61.145 Feature Selection	384	1.851
3.32.697 Formicidae	354	1.707

4.18.204 Distributed Generation	301	1.451
4.61.869 Clustering	295	1.422
4.18.296 Unit Commitment	253	1.22
4.224.599 Project Scheduling	197	0.95
4.224.1307 Process Planning	194	0.935
4.17.282 Image Segmentation	184	0.887
6.153.1330 Water Distribution Systems	172	0.829
4.13.807 Internet Of Things	170	0.82
4,183,486 Traffic Flow	163	0.786
4.13.896 Vehicular Ad Hoc Networks	154	0.743
4.18.472 Voltage Stability	133	0.641
4.47.1111 Software Testing	133	0.641
4.237.651 Preventive Maintenance	131	0.632
4.48.120 Complex Networks	128	0.617
4.84.1965 Cutting Stock Problem	113	0.545
4.61.1302 Intrusion Detection	112	0.54
4.84.1632 Facility Location	112	0.54

When the micro-level distribution of citation topics is evaluated categorically, it is seen that the highest number of publications is in the Optimization and Decision Making Problems category. Particle Swarm Optimization (PSO) (3818 publications) has the highest number of publications among optimization algorithms. This shows that ACO is popularly used in hybrid studies together with PSO. Vehicle Routing Problem with 1435 publications is of great importance in areas such as logistics and supply chain management. Scheduling (1093 publications) reflects research on scheduling and planning problems. Such optimizations play a critical role in manufacturing processes and project management. Subtopics such as Project Scheduling (197 publications) and Process Planning (194 publications) seem to focus on more specific planning and production processes. In the Networks and Distributed Systems category, Wireless Sensor Networks (1167 publications) and Vehicular Ad Hoc Networks (154 publications), Cloud Computing (635 publications). In Machine Learning and Data Analytics category, Feature Selection (384 publications), Clustering (295 publications), in Energy Systems and Distributed Generation category, Distributed Generation (301 publications), Voltage Stability (133 publications) and Unit Commitment (253 publications), in Image Processing and Visual Perception category, Image Segmentation (184 publications). In the Networks and Security category, Intrusion Detection (112 publications) and Complex Networks (128 publications), in the Biology and Environment category, Formicidae (354 publications) and Water Distribution Systems (172 publications), in the Maintenance and Security category, Preventive Maintenance (131 publications) and Software Testing (133 publications), in the Transportation and Traffic category, Traffic Flow (163 publications), Cutting Stock Problem (113 publications) and Facility Location (112 publications).

WoS index distributions

In the analysis, Figure 3 shows the distribution of ACO-related publications according to WoS indexes. This distribution is important to understand through which indexes the relevant literature is published and which fields are dominant.

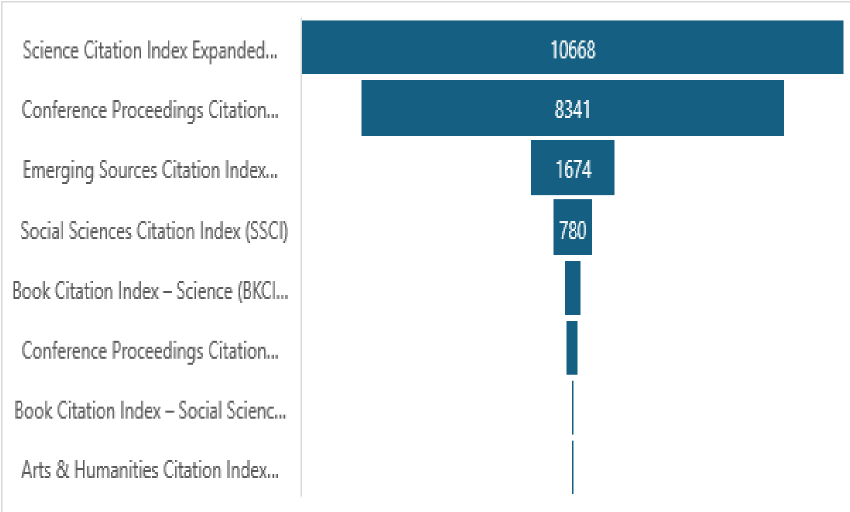


Figure3 : Distribution of publications according to WoS indexes

Figure 3 shows that Science Citation Index Expanded (SCI-EXPANDED), which has the highest number of publications with 10668 publications, includes articles in high quality, high impact factor journals in basic sciences and engineering. Conference Proceedings Citation Index - Science (CPCI-S) With 8341 publications, the conference proceedings index has a large share. Emerging Sources Citation Index (ESCI) has 1674 publications and Social Sciences and Humanities has 780 publications. Other indexes Social Sciences Citation Index (SSCI), Conference Proceedings Citation Index - Social Science & Humanities (CPCI-SSH), Book Citation Index - Science (BKCI-S), Book Citation Index - Social Sciences & Humanities (BKCI-SSH), Arts & Humanities Citation Index (A&HCI).

Distribution of journal names according to publication titles

Table 4 shows the journals in which ACO-related studies were published the most, according to the titles of the publications.

Table4 : Distribution of journal names according to titles of publications

Publication Titles	Record Count	% of 20.740
Lecture Notes in Computer Science	783	3.775
IEEE Access	398	1.919
Applied Soft Computing	337	1.625
Expert Systems with Applications	258	1.244
Ieee Congress on Evolutionary Computation	214	1.032
Lecture Notes in Artificial Intelligence	213	1.027
Applied Mechanics and Materials	187	0.902
Advances In Intelligent Systems and Computing	170	0.82
Computers Industrial Engineering	149	0.718
Communications In Computer and Information Science	148	0.714
Mathematical Problems in Engineering	148	0.714
International Journal of Advanced Manufacturing Technology	142	0.685
Advanced Materials Research	141	0.68
Soft Computing	139	0.67
Studies In Computational Intelligence	132	0.636
Sensors	128	0.617
Journal Of Intelligent Fuzzy Systems	114	0.55
Wireless Personal Communications	114	0.55
Computers Operations Research	106	0.511
International Journal of Production Research	105	0.506
Applied Sciences Basel	102	0.492
Neural Computing Applications	102	0.492
Engineering Applications of Artificial Intelligence	93	0.448
European Journal of Operational Research	87	0.419

Swarm And Evolutionary Computation	84	0.405

When Table 4 is analyzed, it is seen that the highest number of publications is published in Lecture Notes in Computer Science with 783 publications. In other journals, EEE Access 398 publications, Applied Soft Computing 337 publications, Expert Systems with Applications 258 publications, IEEE Congress on Evolutionary Computation 214 publications, Lecture Notes in Artificial Intelligence 213 publications, Applied Mechanics and Materials 187 publications, Advances in Intelligent Systems and Computing: 170 publications, Computers Industrial Engineering 149 publications, Communications in Computer and Information Science 148 publications, Mathematical Problems in Engineering 148 publications, International Journal of Advanced Manufacturing Technology 142 publications, Advanced Materials Research 141 publications, Soft Computing 139 publications, Studies in Computational Intelligence 132 publications, Sensors 128 publications, Journal of Intelligent Fuzzy Systems 114 publications, Wireless Personal Communications 114 publications Computers Operations Research 106 publications International Journal of Production Research: 105 publications, Applied Sciences Basel 102 publications, Neural Computing Applications 102 publications, Engineering Applications of Artificial Intelligence 93 publications, European Journal of Operational Research 87 publications, Swarm and Evolutionary Computation 84 publications.

Language distribution of publications

As a result of the analyses, Figure 5 visualizes the distribution of publications by language of publication. This distribution shows the languages in which publications are published and how the most commonly used languages are used in research.

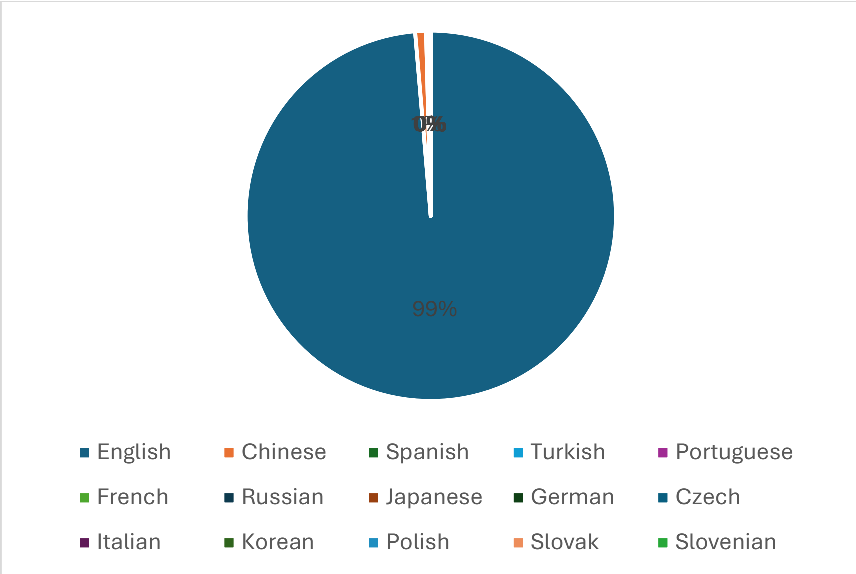


Figure4 : Distribution of publications by language of publication

According to the distribution of publication languages, English: 20,460 publications constitute 98.65 percent of all publications. In other languages, Chinese: 196 publications, Spanish: 26 publications, Turkish: 24 publications, Portuguese: 10 publications.

Publisher distribution of publications

Figure 5 shows the publishing houses with the highest number of publications in the WoS database and their number of publications.

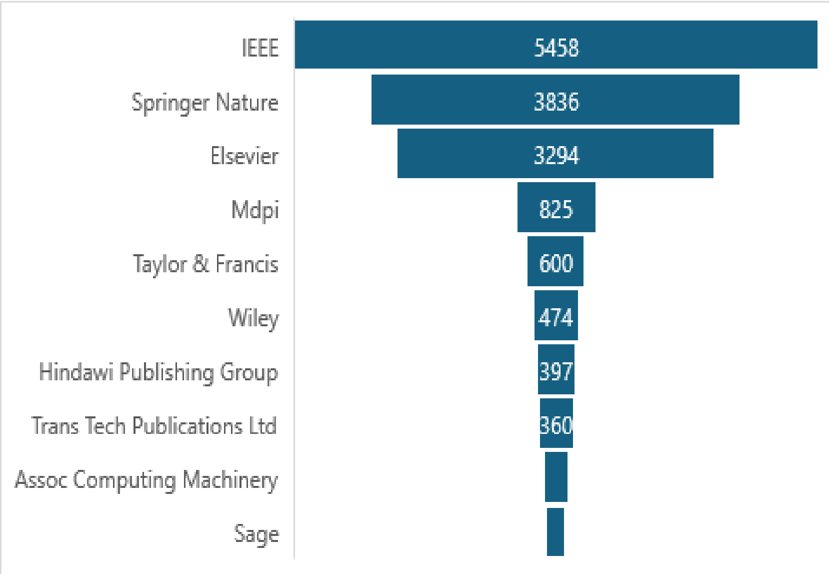


Figure5 : Distribution of publications by publishers

When the publisher distribution of publications is analyzed with Figure 5, IEEE 5458 publications, Springer Nature 3836 publications, Elsevier 3294 publications, MDPI 825 publications, Taylor & Francis 600 publications, Wiley 474 publications, Hindawi Publishing Group 397 publications, Trans Tech Publications Ltd 360 publications, Association for Computing Machinery (ACM) 253 publications, Sage 206 publications are listed.

Co-authorship of authors

As a result of the analyses, Figure 6 presents the links between co-authors of publications. Such analyses provide important information on the intensity of scientific collaborations, the networking between authors, and which researchers collaborate more with each other.

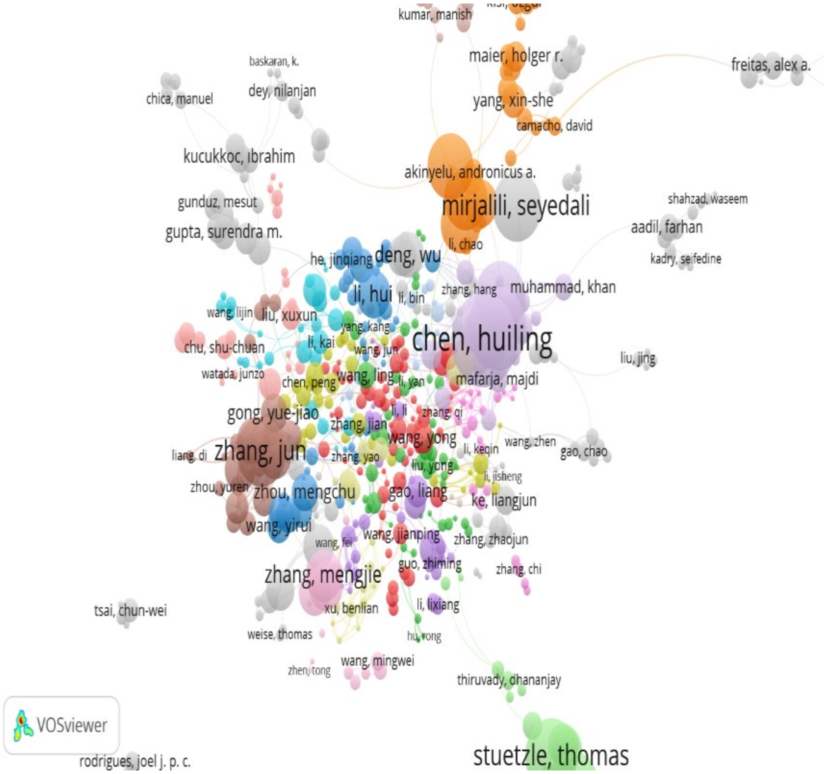


Figure6 : Links between co-authors of publications

Figure 6 shows that Dorigo, Marco has the highest number of citations with 12,829 citations. Although the number of publications is 40, the link strength is 53. Dorigo's high number of citations shows that his research is used by a large number of people. However, the link strength of 53 suggests that it is limited to further academic collaboration or interaction. Stuetzle, Thomas: Ranked second with 8,416 citations and a link strength of 56, suggesting that Stuetzle's publications have also been influential in a wide area, and that his network of interaction (link strength) is slightly higher than Dorigo's. Blum, Christian has 57 publications with 6,404 citations. The link strength is 34, indicating that Blum has a good influence but less direct collaboration. The balance between the number of publications and citations suggests that Blum's research has a wide audience but is disseminated in a more limited academic network. Chen, Huiling has 5,977 citations and 72 publications. With a link strength of 226, he has a very high link strength. Chen has published a large number of publications and these publications seem to have a large network of interactions. This shows that Chen's research is recognized by interacting with a wide academic environment. Heidari, Ali Asghar has 51 publications with 4,773 citations. The link strength is 177, indicating that Heidari has a strong network of academic interaction. The high number of citations and link strength are in line, suggesting that Heidari's research has an impact in a wide circle. Gambardella, LM has 4,396 citations, 13 publications and a link strength of 18.

Citation analysis of organizations with co-authorship

As a result of the analysis, Figure 7 shows the network visualization graph of the institutions according to the number of citations.

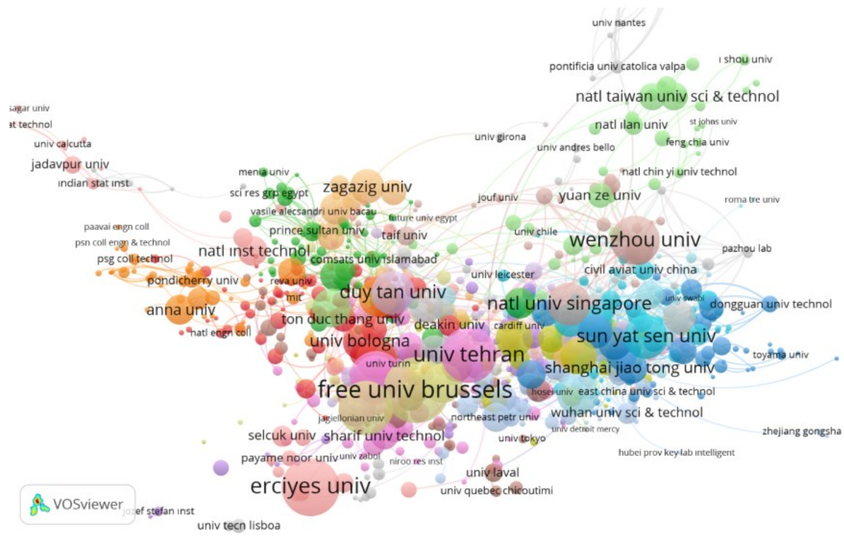


Figure7 : Network visualization graph according to the number of citations of institutions

Figure 7 shows that although Free Univ Brussels has the highest number of citations (10,747), the number of publications (29) is considerably lower than the institutions in the top 5. The link strength of 17 indicates that the work of this institution is relatively less effective in academic collaborations. This shows that they publish few but very effective publications. Erciyes Univ ranks second with 8,201 citations. The number of publications was 49 and the link strength was 16. This shows that Erciyes University is making a significant impact with its work, but remains at a lower level in terms of collaboration or connectivity. Univ Libre Bruxelles stands out with 7,300 citations and 57 publications. The link strength is 32, indicating that the institution is more actively involved in academic collaborations than others. The balanced spread of the number of citations with the link strength shows that this institution is recognized in a wide academic network. Univ Tehran: With 7,570 citations and 161 publications, the institution has a significant publication volume. The link strength of 225 indicates that the institution is part of a wide network of academic cooperation and has a high level of influence. Islamic Azad Univ has the highest number of publications (330) among the top 5 institutions. It has 6,868 citations and has the highest link strength (288). This shows that the institution has created a wide network of academic collaborations by publishing a large number of publications and that these collaborations have had a strong impact on its research.

Keyword analysis

As a result of the analysis, Figure 8 shows the overlay visualization graph according to the number of keywords.

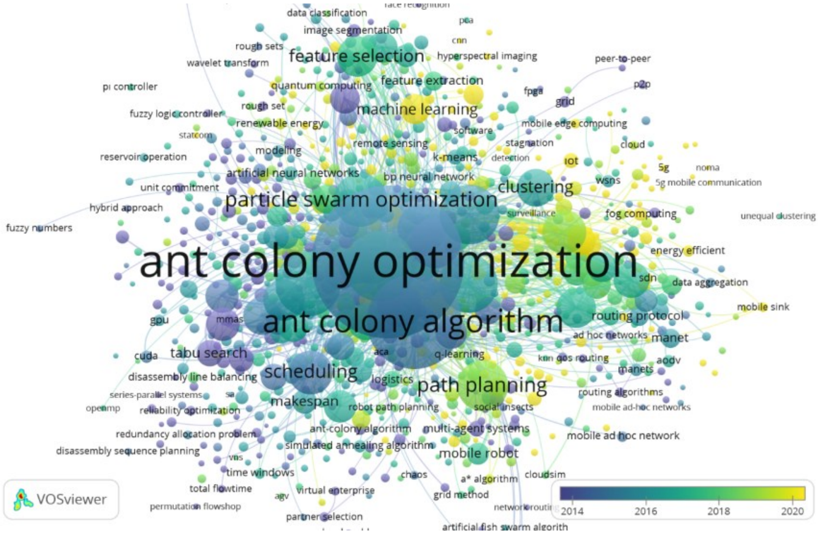


Figure8 : Overlay visualization graph according to the number of keywords

When the data in the graph is evaluated in terms of frequency and link strength, Ant Colony Optimization: 4,692 mentions, total link strength 7,854 and Ant Colony Algorithm: 1,821 mentions, with a total link strength of 2,327, are the most frequent keywords. Other keywords are Optimization: 1,001 occurrences, total link strength 2,723, Genetic Algorithm: 854 mentions, total link strength 2,022, ACO (short for Ant Colony Optimization): 679 mentions, total connection strength 1,431, Particle Swarm Optimization: 579 mentions, total connection strength 1,637, Path Planning: 524 mentions, total connection strength 1,091, Swarm Intelligence: 492 mentions, total connection strength 1,250, Cloud Computing: 396 mentions, total connection strength 1,205, Scheduling: 389 mentions, total connection strength 971.

Citation analysis of publications based on the published work

As a result of the analysis, Figure 9 visualizes the links between the works in which the publications were published and the number of citations these works received.

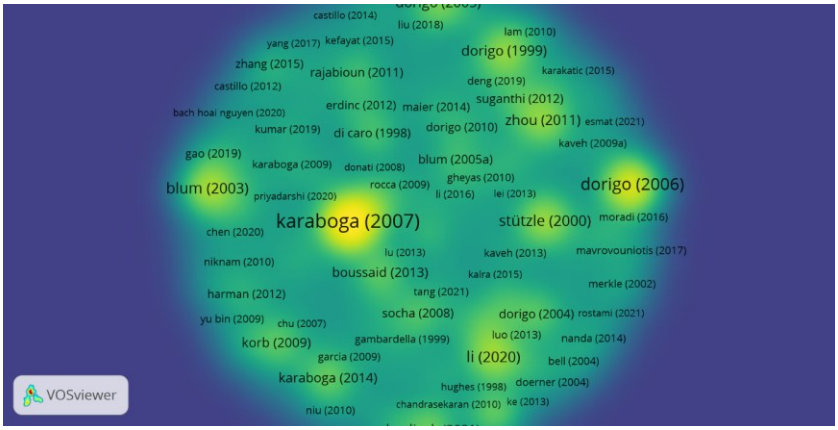


Figure9 : Citation analysis of publications on the basis of the published work

According to the citation analysis of the works published with Figure 9, it is seen that the work by (Karaboga & Basturk, 2007) ranks first with 5111 citations. A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm focuses on the Artificial Bee Colony (ABC) algorithm and compares it with other popular meta-heuristic algorithms. (Dorigo et al., 2006) "Ant colony optimization", which introduces colony optimization and examines its most remarkable applications, ranks second with 3151 citations. (Blum & Roli, 2003) "Metaheuristics in combinatorial optimization: Overview and conceptual comparison" by ranks third with 2066 citations. In this study, the advantages and disadvantages of different metaheuristics are outlined and the importance of hybridization and integration of metaheuristics is emphasized. (The study titled MAX - MIN Ant System by Stützle & Hoos, 2000) ranks fourth with 1846 citations. In this study, it is seen that the performance improvements of ACO were made and successful experimental results were obtained on the traveling salesman problem and the quadratic assignment problem. (Li et al., 2020) by "Slime mould algorithm: A new method for stochastic optimization". This study ranks fifth with 1829 citations. In this study, the researchers compared their newly developed algorithm with ACO and other popular meta-heuristic algorithms and shared the successful results of their newly developed algorithm.

Link between citations in the journals in which the publication was made

As a result of the analysis, Figure 10 shows the link between the number of publications and citations in the journals in which the publications are published.

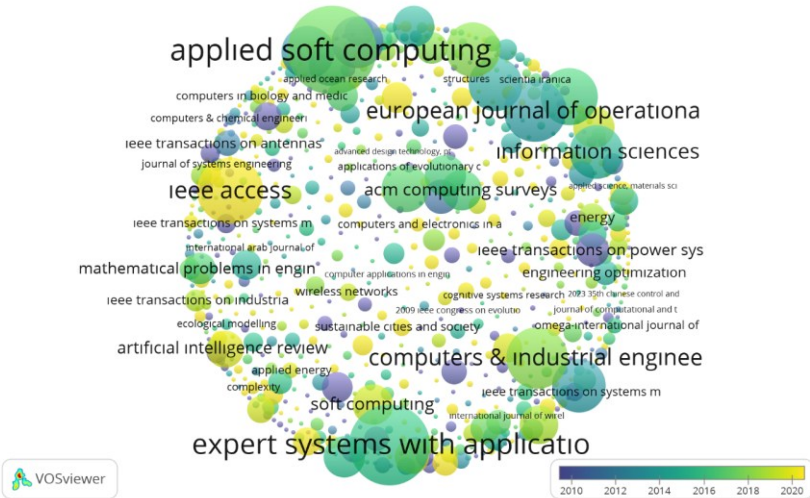


Figure10 : The link between the number of publications and citations in the journals in which the publications are published

When the journals publishing in ACO literature are analyzed in terms of number of citations and number of documents, Applied

Soft Computing ranks first: 337, Citations: 15,132. This journal receives many citations in the field of artificial intelligence and optimization. Expert Systems with Applications ranks second with 12,508 citations in 258 documents. Focusing on expert systems and applications, this journal covers the use of meta-heuristic algorithms for practical engineering problems. IEEE Access ranks third with 8422 citations in 398 papers. Covering a wide range of engineering and technology, IEEE Access is frequently cited due to its high access rate. European Journal of Operational Research ranks fourth with 7,439 citations in 87 documents. This journal includes theoretical and applied studies on operational research and publishes studies on optimization algorithms such as ACO. Computers & Industrial Engineering ranked fifth with a total of 7346 citations in 149 documents. Focusing on industrial engineering and computer-aided applications, this journal has made a significant contribution to the ACO literature.

3. Results

According to the WoS category distributions, computer science categories are leading in the total number of publications. This shows that the ACO algorithm is primarily developed and applied to computer science-based problems. Engineering categories emphasize that the ACO algorithm can be easily adapted to different engineering applications and has a multidisciplinary nature. Energy and Environmental Sciences (551 and 403) indicate the increasing use of the ACO algorithm for critical problems such as sustainability and energy efficiency.

The distribution of publications by year shows that an increase started in the late 1990s and gained momentum rapidly after the 2010s. Especially after 2017, scientific production is thought to have accelerated further with the impact of digitalization and global collaborations. This situation shows that research and publication processes have become more accessible and academic studies are spreading at an increasing speed.

The distribution of publications by document type shows that articles and conference proceedings have a large share, in other words, most of the academic production takes place in these two types. Other document types have a much smaller share and are usually published for specific needs.

According to the meso-distribution data of the citation topics of the publications, it is revealed that technology, sustainability, health sciences and education fields come to the forefront in the research world. Topics such as artificial intelligence, machine learning, robotics and telecommunications show that technology-oriented research is rapidly increasing. At the same time, it is shown that there is intense research activity on global issues such as environmental sustainability and climate change. This shows that the scientific community is focusing on both future technological innovations and solutions to social and environmental problems.

The micro distribution data of the citation topics of the publications show that in-depth studies have been conducted on more specific and technical research topics on ACO. Optimization problems, especially PSO, transportation and scheduling, are critical for increasing efficiency in manufacturing and logistics. Networks and security issues reveal that research on wireless sensor networks, cloud computing and cybersecurity is increasing in a world of rapidly increasing digitalization.

According to WoS index distributions, SCI-EXPANDED and CPCI-S indexes cover a large proportion of scientific research and show that studies in these fields receive more citations. Other indexes are SSCI, CPCI-SSH, BKCI-S, BKCI-SSH and ESCI.

According to the distribution of publication languages, English has a very large proportion of publications, while the number of publications in other languages is more limited.

According to the publisher distribution data of the publications, IEEE, Springer Nature and Elsevier stand out as the publishers with the highest number of publications, while the number of publications of other publishers lags behind these three publishers.

While authors such as Dorigo, Marco and Stuetzle, Thomas have high citation counts and strong academic influence, authors such as Chen, Huiling and Heidari, Ali Asghar have a stronger academic interaction network. Chen, Huiling has the highest link strength, indicating that she has a wide interaction and collaboration network. Gambardella, LM, on the other hand, can be said to have a narrower sphere of influence compared to other authors with low number of publications and link strength.

According to the results of the citation analysis data of co-authored institutions, Free Univ Brussels has the highest number of citations, but the small number of publications and low link strength indicate a narrow but influential field of study. Islamic Azad Univ has the largest publication volume and the strongest network. This shows that the institution is a leader in global academic collaborations. Univ Tehran stands out with its high connectivity and balanced citation-publication performance, while Erciyes Univ and Univ Libre Bruxelles have a more limited impact and connectivity.

Ant Colony Optimization and its derivatives (ACO, Ant Colony Algorithm, Ant Colony Optimization Algorithm) dominate the keywords. Other meta-heuristic algorithms such as Genetic Algorithm and Particle Swarm Optimization are used in combination with ACO. Terms such as Optimization and Cloud Computing indicate that the methods are used extensively both theoretically and practically.

The results of the citation analysis of the publications on the basis of the published work reveal important contributions made in different time periods in the development process of the ACO algorithm. In particular, (Karaboga & Basturk, 2007) and (Dorigo et al., 2006) are the main works referenced in algorithm development and implementation.

According to the results of the relationship between the number of publications and the number of citations in the journals, it was concluded that the highest number of citations belonged to the journal Applied Soft Computing with 15,132 citations in 337 documents.

References

Akkaya, A., Közkurt C. & Durgut R. (2023). A study on meta-heuristic algorithms used for problem solving in recent years. *Pioneer and Contemporary Studies in* Engineering 18, 327-352.

Ahmad, S. R., Yusop, N. M. M., Bakar, A. A., & Yaakub, M. R. (2017). Statistical analysis for validating ACO-KNN algorithm as feature selection in sentiment analysis. *AIP Conference Proceedings*,*1891* (1), 020018. https://doi.org/10.1063/1.5005351

Aria, M., & Cuccurullo, C. (2017). *bibliometrix*: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*,*11* (4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

Blum, C., & Roli, A. (2003). Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Comput. Surv.*,*35* (3), 268-308. https://doi.org/10.1145/937503.937505

Chaudhari, K., & Thakkar, A. (2019). Travelling Salesman Problem: An Empirical Comparison Between ACO, PSO, ABC, FA and GA. In N. R. Shetty, L. M. Patnaik, H. C. Nagaraj, P. N. Hamsavath, & N. Nalini (Eds.), *Emerging Research in Computing, Information, Communication and Applications* (ss. 397-405). Springer. https://doi.org/10.1007/978-981-13-6001-5_32

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*,*133* , 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070

M. Dorigo and L. M. Gambardella, "Ant colony system: A cooperative learning approach to the traveling salesman problem," IEEE Transactions on Evolutionary Computation, vol. 1, no. 1, pp. 53-66, Apr. 1997.

Dorigo, M., & Stützle, T. (2004). *Ant Colony Optimization*. Cambridge, MA: MIT Press.

Dorigo, M., Birattari, M., & Stutzle, T. (2006). Ant colony optimization. *IEEE Computational Intelligence Magazine*,*1* (4), 28-39. IEEE Computational Intelligence Magazine. https://doi.org/10.1109/MCI.2006.329691

Durgut, R. (2021). Improved binary artificial bee colony algorithm. *Frontiers of Information Technology & Electronic Engineering*,*22* (8), 1080-1091. https://doi.org/10.1631/FITEE.2000239

Eid, M., Alassery, F., Ibrahim, A., & Saber, M. (2022). Metaheuristic Optimization Algorithm for Signals Classification of Electroencephalography Channels. *Computers, Materials & Continua*,*71* (3), 4627-4641. https://doi.org/10.32604/cmc.2022.024043

Emami Skardi, M. J., Afshar, A., Saadatpour, M., & Sandoval Solis, S. (2015). Hybrid ACO-ANN-Based Multi-objective Simulation-Optimization Model for Pollutant Load Control at Basin Scale. *Environmental Modeling & Assessment*,*20* (1), 29-39. https://doi.org/10.1007/s10666-014-9413-7

Ganji, M. F., & Abadeh, M. S. (2010). Parallel fuzzy rule learning using an ACO-based algorithm for medical data mining. *2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)*, 573-581. https://doi.org/10.1109/BICTA.2010.5645189

Karaboga, D., & Basturk, B. (2007). A powerful and efficient algorithm for numerical function optimization: Artificial bee colony (ABC) algorithm. *Journal of Global Optimization*,*39* (3), 459-471. https://doi.org/10.1007/s10898-007-9149-x

Kaveh, A., & Talatahari, S. (2010). An improved ant colony optimization for constrained engineering design problems. *Engineering Computations*,*27* (1), 155-182. https://doi.org/10.1108/02644401011008577

Li, S., Chen, H., Wang, M., Heidari, A. A., & Mirjalili, S. (2020). Slime mould algorithm: A new method for stochastic optimization. *Future Generation Computer Systems*,*111* , 300-323. https://doi.org/10.1016/j.future.2020.03.055

López-Ibáñez, M., & Stützle, T. (2010, September). Automatic configuration of multi-objective ACO algorithms. In *International Conference on Swarm Intelligence* (pp. 95-106). Berlin, Heidelberg: Springer Berlin Heidelberg.

Ma, S., Yang, R., Deng, X., Ling, X., Zhang, X., Zhou, F., Li, S., & Kwan Ng, D. W. (2022). Spectral and Energy Efficiency of ACO-OFDM in Visible Light Communication Systems. *IEEE Transactions on Wireless Communications*,*21* (4), 2147-2161. IEEE Transactions on Wireless Communications. https://doi.org/10.1109/TWC.2021.3109735

Madridano, Á., Al-Kaff, A., Martín, D., & de la Escalera, A. (2021). Trajectory planning for multi-robot systems: Methods and applications. *Expert Systems with Applications*,*173* , 114660. https://doi.org/10.1016/j.eswa.2021.114660

Mardoude, Y., Rahali, A., & Hilali, A. (2024). Maximizing Wind Energy Efficiency Through ACO and ANN-Based MPPT

Algorithms. In S. Motahhir & B. Bossoufi (Eds.), *Digital Technologies and Applications* (pp. 363-373). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-68653-5_35

Mavrovouniotis, M., & Yang, S. (2015). Ant algorithms with immigrants schemes for the dynamic vehicle routing problem. *Information Sciences*, *294*, 456-477.

Ming, Y., Li, Y., Zhang, Z., & Yan, W. (2021). A Survey of Path Planning Algorithms for Autonomous Vehicles. *SAE International Journal of Commercial Vehicles*,*14*(1), Article 02-14-01-0007. https://doi.org/10.4271/02-14-01-0007

Price, D. J. D. S. (2019). *Little Science, Big Science*. Columbia University Press. https://doi.org/10.7312/pric91844

Revanna, J. K. C., & Al-Nakash, N. Y. B. (2023). Metaheuristic link prediction (MLP) using AI based ACO-GA optimization model for solving vehicle routing problem. *International Journal of Information Technology*,*15* (7), 3425-3439. https://doi.org/10.1007/s41870-023-01378-5

Safavat, S., & Rawat, D. B. (2021). On the Elliptic Curve Cryptography for Privacy-Aware Secure ACO-AODV Routing in Intent-Based Internet of Vehicles for Smart Cities. *IEEE Transactions on Intelligent Transportation Systems*.*22* (8), 5050-5059. IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/TITS.2020.3008361

Sandini, G., Sciutti, A., & Morasso, P. (2024). Artificial cognition vs. Artificial intelligence for next-generation autonomous robotic agents. *Frontiers in Computational Neuroscience*,*18* . https://doi.org/10.3389/fncom.2024.1349408

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*,*24* (4), 265-269. https://doi.org/10.1002/asi.4630240406

Socha, K., & Dorigo, M. (2008). Ant colony optimization for continuous domains. *European journal of operational research*, *185*(3), 1155-1173.

Stützle, T., & Hoos, H. H. (2000). MAX-MIN Ant System. *Future Generation Computer Systems*,*16* (8), 889-914. https://doi.org/10.1016/S0167-739X(00)00043-1

Vanderplaats, G. N. (1988). Multidiscipline Design Optimization. *Applied Mechanics Reviews*,*41* (6), 257-262. https://doi.org/10.1115/1.3151897

Wang, C., Zhang, G., Xu, H., & Chen, H. (2016). An ACO-based Link Load-Balancing Algorithm in SDN. *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, 214-218. https://doi.org/10.1109/CCBD.2016.050

Zhang, C., Cai, Y., Hu, P., Quan, P., & Song, W. (2022). Logistics Distribution Route Optimization Using Hybrid Ant Colony Optimization Algorithm. In X. Zhao, S. Yang, X. Wang, & J. Li (Eds.), *Web Information Systems and Applications* (pp.. 510-517). Springer International Publishing. https://doi.org/10.1007/978-3-031-20309-1_45

CHAPTER IV

A Prediagnostic Tool for PCOS: Integrating Machine Learning and Web-Based Solutions

Hakan YILMAZ[1]
Mehmet ÖZDEM[2]

1. Introduction

Polycystic Ovary Syndrome (PCOS) is a prevalent endocrine disorder that primarily affects women of reproductive age, characterized by a complex interplay of reproductive, metabolic, and psychological disturbances. The condition is defined by a combination of clinical features, including oligo- or amenorrhea, hyperandrogenism, and polycystic ovaries, as established by the Rotterdam criteria (Kurniawati, Pramono, Hidayat, & Mahati, 2024). The etiology of PCOS is multifactorial, involving genetic predispositions, environmental factors, and hormonal imbalances, particularly concerning insulin resistance and hyperandrogenemia (Liu & Zhu, 2024; Zhao, Zhang, Cheng, Nie, & He, 2023).

[1] Doç. Dr., Karabük Üniversitesi, Mühendislik Fakültesi, Tıp Mühendisliği Bölümü, Karabük/Türkiye, Orcid: 0000-0002-8553-388X, hakanyilmaz@karabuk.edu.tr
[2] Dr., Türk Telekom, Ankara/Türkiye, Orcid: 0000-0002-2901-2342, mehmet.ozdem@turktelekom.com.tr

Recent studies have highlighted the significant role of insulin resistance in the pathogenesis of PCOS, which exacerbates metabolic dysfunction and increases the risk of developing type 2 diabetes and cardiovascular diseases (Ganie et al., 2019; Zhao et al., 2023). Furthermore, the hormonal dysregulation associated with PCOS impacts various organ systems, leading to a spectrum of clinical manifestations, including hirsutism, acne, and infertility (Emanuel et al., 2022; Glintborg & Andersen, 2010). The complexity of PCOS is further underscored by its association with obesity, which can complicate the clinical picture and influence treatment outcomes (G. Usha Kiran et al., 2023; Khamoshina, Artemenko, Bayramova, Ryabova, & Orazov, 2022).

Management strategies for PCOS are diverse and may include lifestyle modifications, pharmacological interventions, and assisted reproductive technologies. Medications such as metformin and hormonal contraceptives are commonly employed to address insulin resistance and regulate menstrual cycles (Aydın, Dayan, & Güngör, 2023; M. L. Misso & Teede, 2015). Additionally, surgical options like bariatric surgery have shown promise in improving metabolic parameters and PCOS-specific characteristics in obese patients (Ghobrial, Ott, Steininger, Dewailly, & Prager, 2023; Khamoshina et al., 2022). However, the need for individualized treatment approaches remains critical, given the variability in symptoms and underlying pathophysiology among affected individuals (Aydın et al., 2023; Kumari & Jha, 2024). In summary, PCOS is a multifaceted disorder that necessitates a comprehensive understanding of its pathophysiology, clinical manifestations, and management strategies. Ongoing research is essential to unravel the complexities of PCOS and to develop effective therapeutic interventions tailored to the unique needs of each patient (Aishwarya, Sunitha, & Chandy, 2023; Hiam et al., 2019).

The diagnostic process for PCOS has evolved, with recent guidelines emphasizing a multidisciplinary approach that integrates clinical expertise, patient preferences, and the latest evidence-based practices (Al Wattar et al., 2021; M. Misso et al., 2018). However, discrepancies in diagnostic criteria and the subjective nature of symptom reporting often lead to delays in diagnosis and misdiagnosis, particularly among adolescents (Peña, Teede, Hewawasam, Hull, & Gibson-Helm, 2022; Trent & Gordon, 2020). As such, there is a pressing need for standardized diagnostic protocols that can enhance the accuracy and timeliness of PCOS diagnoses, thereby improving patient care and management (Varanasi et al., 2018).

The application of machine learning (ML) in the diagnosis of Polycystic Ovary Syndrome has garnered significant attention in recent years due to its potential to enhance diagnostic accuracy and facilitate early detection. Traditional diagnostic methods often rely on clinical assessments and biochemical tests, which can be subjective and time-consuming. In contrast, machine learning techniques offer a systematic approach to analyze large datasets, thereby improving diagnostic precision and efficiency.

Recent studies have demonstrated the effectiveness of various machine learning algorithms in diagnosing PCOS. For instance, a comprehensive analysis of multiple algorithms such as Support Vector Machines (SVM), Random Forest, and deep learning architectures has shown promising results in identifying PCOS based on clinical and biochemical features (Kadam, 2024). The integration of multimodal data, including ultrasound images and patient history, has further enhanced diagnostic capabilities, allowing for a more holistic view of the patient's condition (Nethra Sai, Sakthivel, Prakash, Vishnukumar, & Dugki, 2024). This multimodal approach is crucial given the heterogeneous nature of PCOS, which manifests differently across individuals.

Deep learning, particularly through Convolutional Neural Networks (CNNs), has emerged as a powerful tool in the automated identification of PCOS. Studies have reported the successful application of CNNs in analyzing ultrasound images to detect PCOS with high accuracy (Sunil, Swathi, Chidvi Chaturya, Sowjanya, & Gowreeswari, 2024). The ability of deep learning models to learn complex patterns from large datasets makes them particularly suited for medical diagnostics, where variations in data can be subtle yet clinically significant. For example, a recent study achieved an impressive classification accuracy of 95% using a custom deep learning model specifically designed for PCOS detection (Sowmiya et al., 2024).

Moreover, ensemble learning methods, which combine multiple algorithms to improve predictive performance, have also been explored in the context of PCOS diagnosis. Research has indicated that hybrid models, such as those combining XGBoost and Random Forest, can significantly enhance detection rates by leveraging the strengths of different algorithms (A, Jain, Namratha, S, & Lakshmi, 2023; Kumar & Varadarajan, 2024). These approaches not only improve accuracy but also provide insights into the most relevant features associated with PCOS, aiding in the development of targeted interventions.

The use of electronic health records (EHR) in conjunction with machine learning algorithms has further expanded the potential for early detection of PCOS. Studies utilizing large-scale EHR data have identified key predictors of PCOS, such as hormonal levels and obesity, which can guide clinicians in risk assessment and management strategies (Zad et al., 2023, 2024). This integration of machine learning with EHR data underscores the importance of utilizing comprehensive datasets to enhance diagnostic accuracy and facilitate timely interventions.

In conclusion, the application of machine learning in the diagnosis of PCOS represents a significant advancement in medical technology. By leveraging various algorithms and integrating multimodal data, researchers are developing robust diagnostic tools that promise to improve the accuracy and efficiency of PCOS detection. Continued research in this area is essential to validate these models across diverse populations and clinical settings, ultimately leading to better health outcomes for women affected by this condition.

The aim of this study is to develop an accurate, efficient, and user-friendly predictive system for Polycystic Ovary Syndrome (PCOS) by leveraging advanced data preprocessing techniques and machine learning algorithms.

2. Materials And Methods

2.1. Dataset Used

The dataset used in this study was obtained from a platform that publishes publicly available datasets (Kottarathil, 2020). The dataset, published by Kottarathil, consists of physical and clinical indicators related to PCOS. It includes a total of 42 input features such as blood type, resting heart rate, height-weight information, age, marital status, pregnancy status, length of menstrual cycle, FSH level, TSH level, and others. In contrast to these input features, there is an output label indicating whether the subject has PCOS or not. The dataset consists of 43 columns and 541 rows in total.

2.2. Programming Environment

The dataset was processed, machine learning algorithms were applied, model metrics were extracted, and suitable methods were selected using a computer equipped with an Intel i7 processor and 16 GB of RAM. These tasks were performed utilizing the Python programming language. Additionally, the webpage design was implemented using HTML, CSS, and JavaScript.

2.3. Metrics Used in Performance Evaluation

In machine learning, particularly in classification tasks, evaluating model performance using various metrics is crucial. These metrics provide essential insights into a model's effectiveness and facilitate the comparison of different models. Prominent metrics include accuracy, sensitivity, specificity, the F1 score, the confusion matrix, and the Receiver Operating Characteristic (ROC) curve.

The confusion matrix serves as a foundational tool for evaluating the performance of classification algorithms. It is a tabular representation that contrasts the model's predicted classifications with the actual classifications, offering a detailed analysis of performance. The matrix comprises four components:

- True Positives (TP): Cases correctly predicted as positive.

- True Negatives (TN): Cases correctly predicted as negative.
- False Positives (FP): Cases incorrectly predicted as positive (Type I error).
- False Negatives (FN): Cases incorrectly predicted as negative (Type II error).

Accuracy indicates the proportion of correct predictions (true positives and true negatives) to the total number of instances, calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity (or recall) measures the percentage of actual positive instances correctly identified by the model. It is especially critical in contexts like medical diagnostics, where failing to detect a positive instance (e.g., a disease) could have severe consequences. The formula for sensitivity is:

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity quantifies the percentage of actual negative cases accurately classified by the model. It holds importance in scenarios where false positives could lead to unnecessary interventions or anxiety. The formula for specificity is:

$$Specificity = \frac{TN}{TN + FP}$$

The F_1 score represents the harmonic mean of precision and recall, offering a balanced evaluation metric, particularly for imbalanced datasets. Ranging from 0 to 1, a higher F_1 score indicates better precision and recall. It is calculated as:

$$\mathrm{F}_1\ Score = \frac{2 * TP}{2 *\ TP + FP + FN}$$

The ROC curve provides a visual depiction of model performance across various classification thresholds by plotting the True Positive Rate (sensitivity) against the False Positive Rate (1 -

specificity). The Area Under the Curve (AUC) summarizes the overall performance of the model, with higher AUC values indicating greater effectiveness (Yılmaz, 2024).

3. Experimental Study

3.1. Data Preprocessing

The study plans to create a web page that will predict the PCOS status of individuals based on their physical and clinical indicators. For this, the user needs to input 42 data points. This could be time-consuming and potentially confusing. To reduce these drawbacks, various preprocessing techniques were applied to reduce the number of input features. For the feature reduction process, Recursive Feature Elimination (RFE) and Feature Importance (FI) methods were applied to the data.

Recursive Feature Elimination (RFE) is a widely recognized technique in the domain of feature selection, particularly within machine learning applications. RFE operates by recursively removing the least important features from a dataset, thereby enhancing the model's performance by focusing on the most relevant attributes. This method has gained traction due to its effectiveness in improving classification accuracy and reducing overfitting, as it systematically narrows down the feature set to those that contribute the most to the predictive power of the model (Jeon & Oh, 2020; Priyatno, 2024a). The algorithm's iterative nature allows it to evaluate the importance of features based on their contribution to the model's performance, making it a valuable tool in various fields, including bioinformatics, finance, and healthcare (A. R. Gupta & Agrawal, 2021; Napa, 2024; Priyatno, 2024b).

Feature Importance (FI), on the other hand, refers to the techniques used to assign a score to each feature based on its relevance to the predictive model. This scoring can be derived from various algorithms, such as decision trees, random forests, or support vector machines, which inherently provide measures of feature importance (R. K. Gupta, Kleinjans, & Caiment, 2021; Kusnawi, 2024). The integration of FI with RFE enhances the feature selection process by ensuring that the features retained are not only relevant but also contribute significantly to the model's predictive capabilities. For instance, studies have shown that combining RFE with feature importance metrics can lead to improved model accuracy and interpretability, as it allows practitioners to understand which features are driving predictions (Priyatno, 2023; Thangapriya, 2024).

After applying RFE and FI to the dataset, the features were ranked in descending order based on their influence on the result for each method. The top 20 features for each method were selected, and the union of these features resulted in a total of 24 selected features. The remaining features were removed from the dataset. As a result, the final dataset consists of 24 input features and 1 output feature, totalling 25 columns. Among these, 16 features are numeric, and 8 are categorical. Outlier tests were applied to the columns containing numeric data, and the extreme values were removed from the dataset. After removing 7 rows, the total number of rows is 534. The statistical summary of the numeric data is presented in Table 1. Additionally, the distribution graphs of categorical data are given in Figure 1a and Figure 1b, and the distribution graphs of numerical data are given in Figure 2a, Figure 2b, Figure 2c and Figure 2d.

Table 1: Statistical summary of numeric data.

Numerical features	**Min.**	**Median**	**Mean**	**Std**	**Max**
Duration of marriage (years)	0	7	7.70	4.78	30
FSH/LH	0.23	2.17	3.80	5.41	61.88
Follicle number (left)	0	5	6.11	4.25	22
LH(mIU/mL)	0.02	2.245	2.73	2.30	14.69

BMI	12.42	24.23	24.29	4.04	38.90
Cycle length (days)	2	5	4.95	1.48	12
Endometrium (mm)	0	8.5	8.50	2.14	18
AMH (ng/mL)	0.1	3.7	5.51	5.29	32
Age (years)	20	31	31.44	5.39	48
PRG (ng/mL)	0.05	0.32	0.41	0.39	6.39
TSH (mIU/L)	0.04	2.26	3.00	3.78	65
Average follicle size (R) (mm)	0	16	15.44	3.33	24
Follicle number (right)	0	6	6.60	4.44	20
FSH (mIU/mL)	0.21	4.84	5.16	3.68	60.37
Hip (cm)	66.0	96.5	96.45	10.10	121.9
Pulse rate (bpm)	13	72	73.25	4.44	82

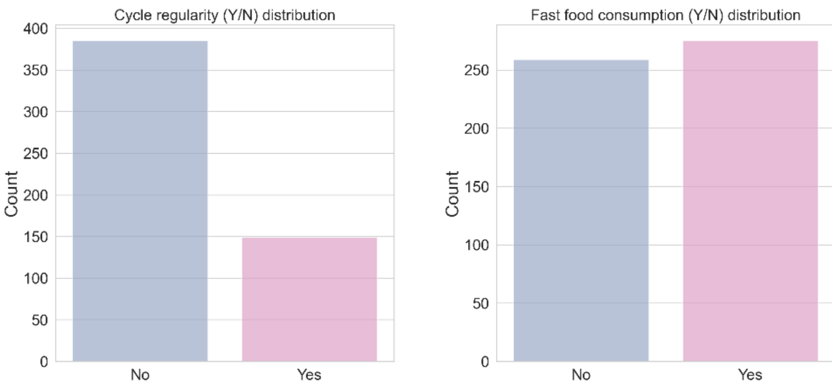


Figure 1a: Distributions of categorical data.

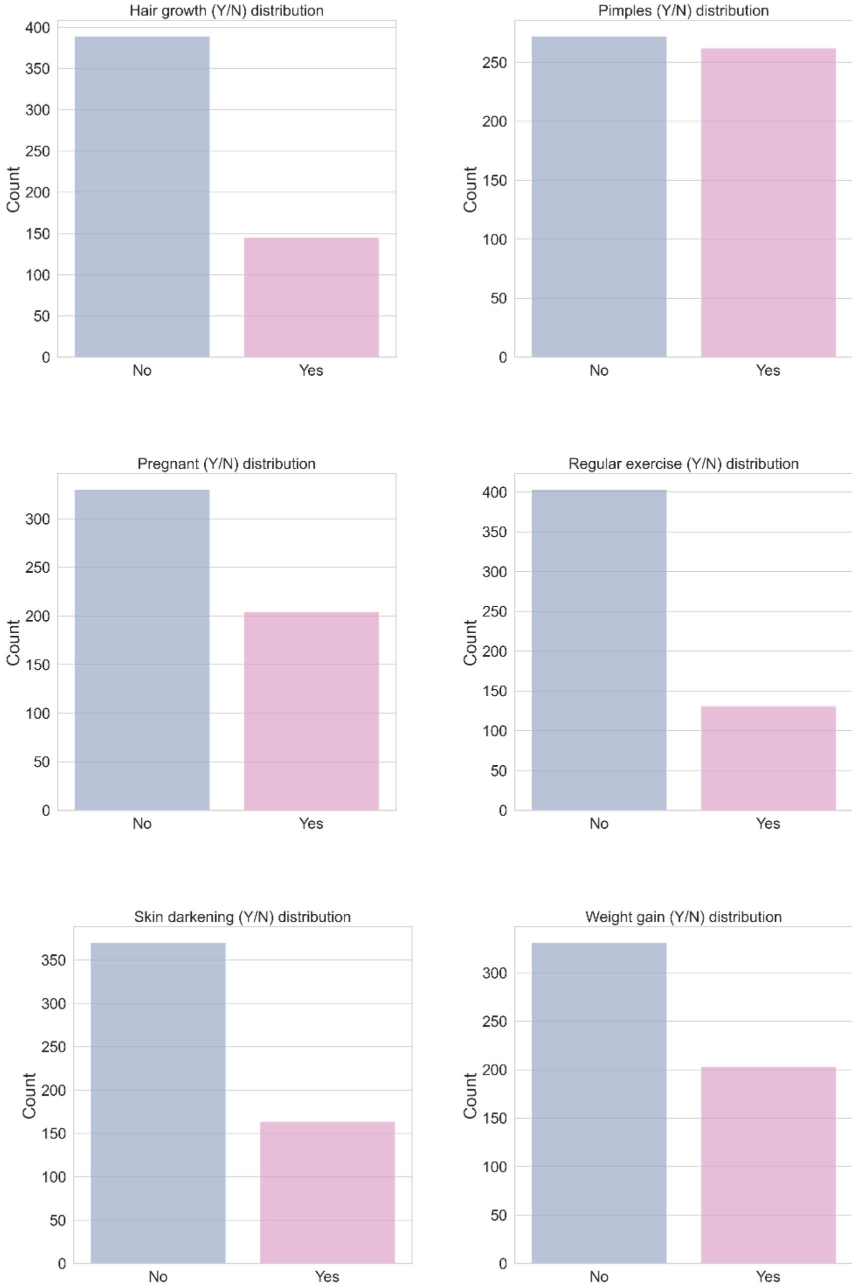


Figure 1b: Distributions of categorical data.

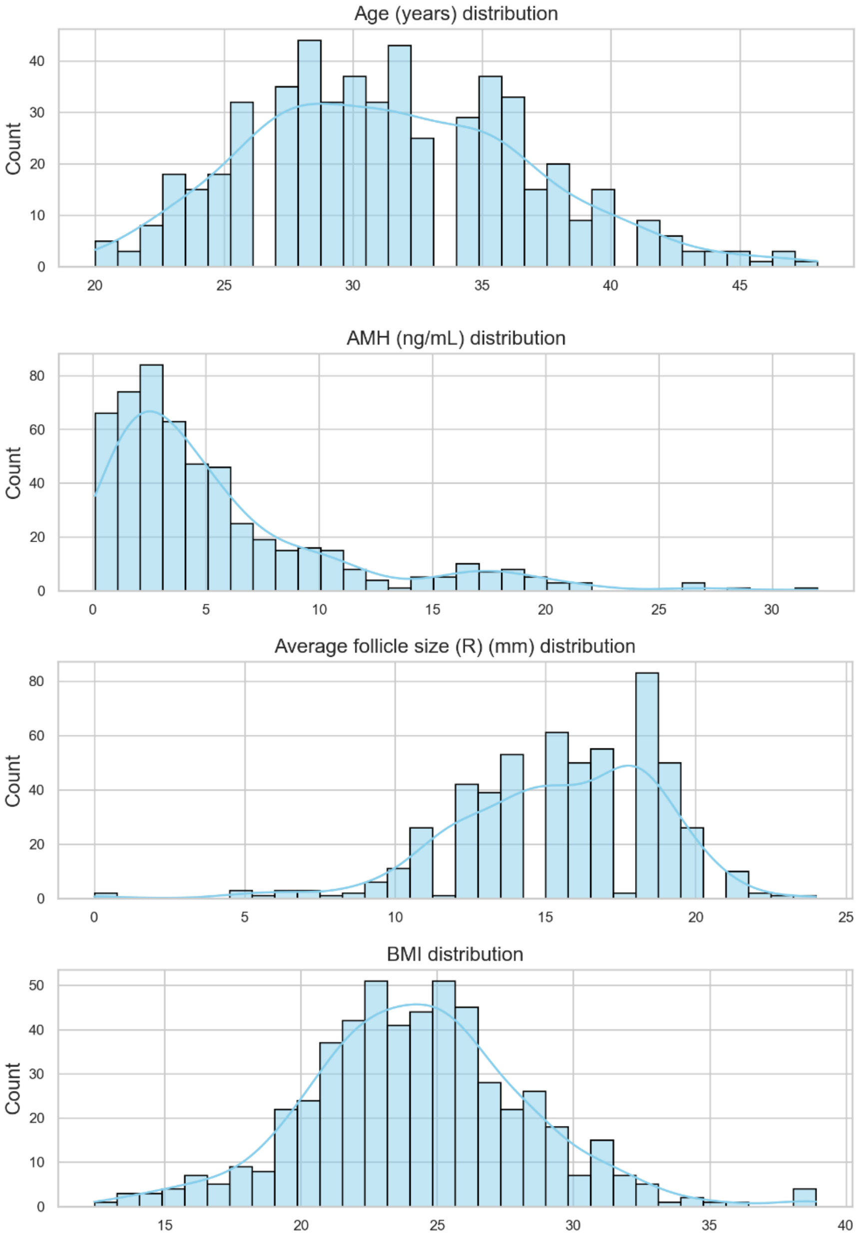


Figure 2a: Distributions of numeric data.

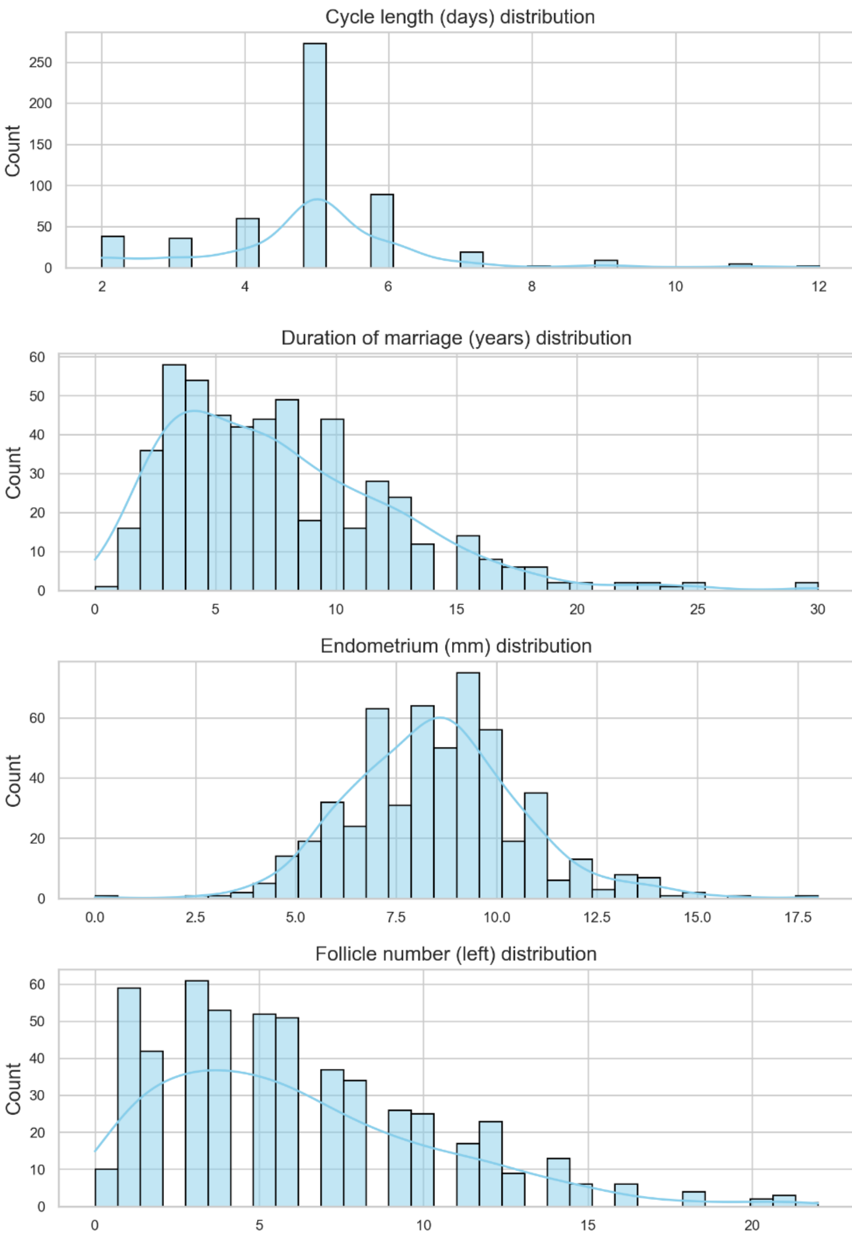


Figure 2b: Distributions of numeric data.

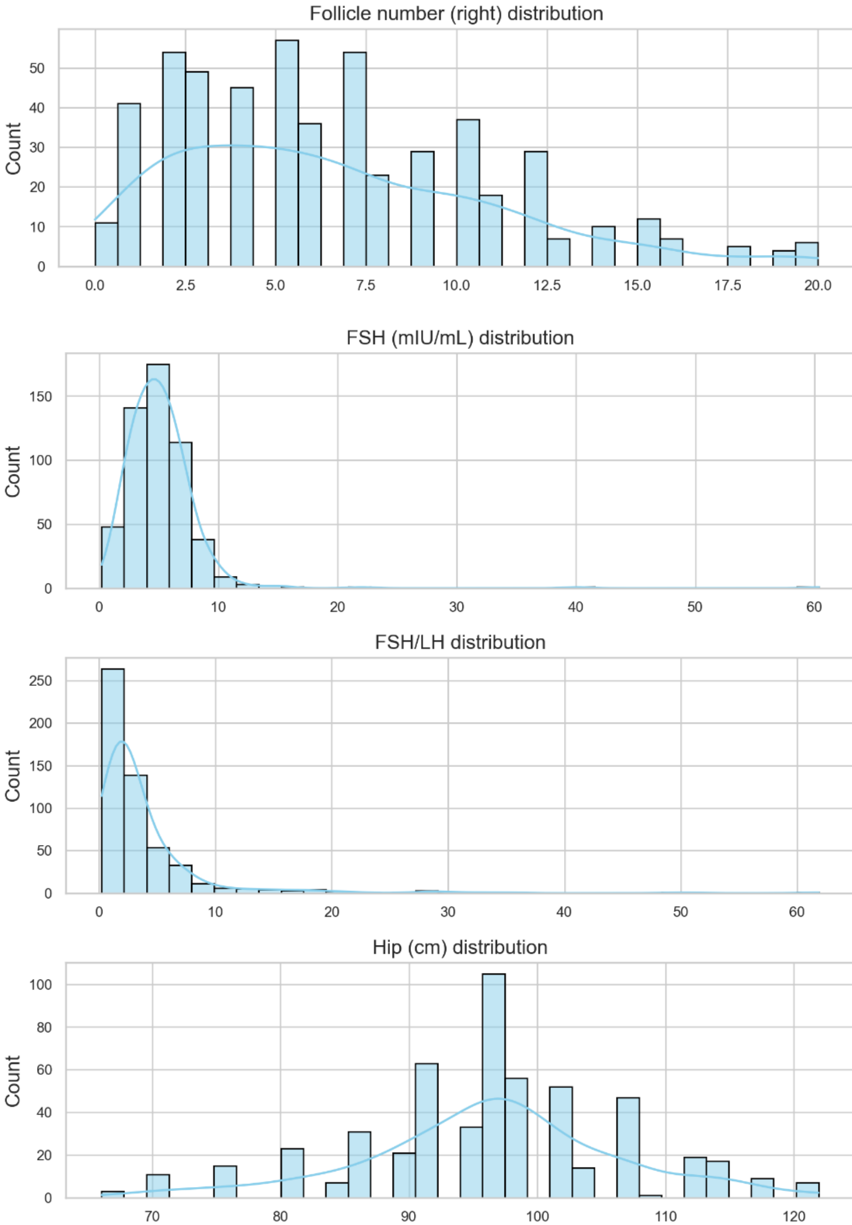


Figure 2c: Distributions of numeric data.

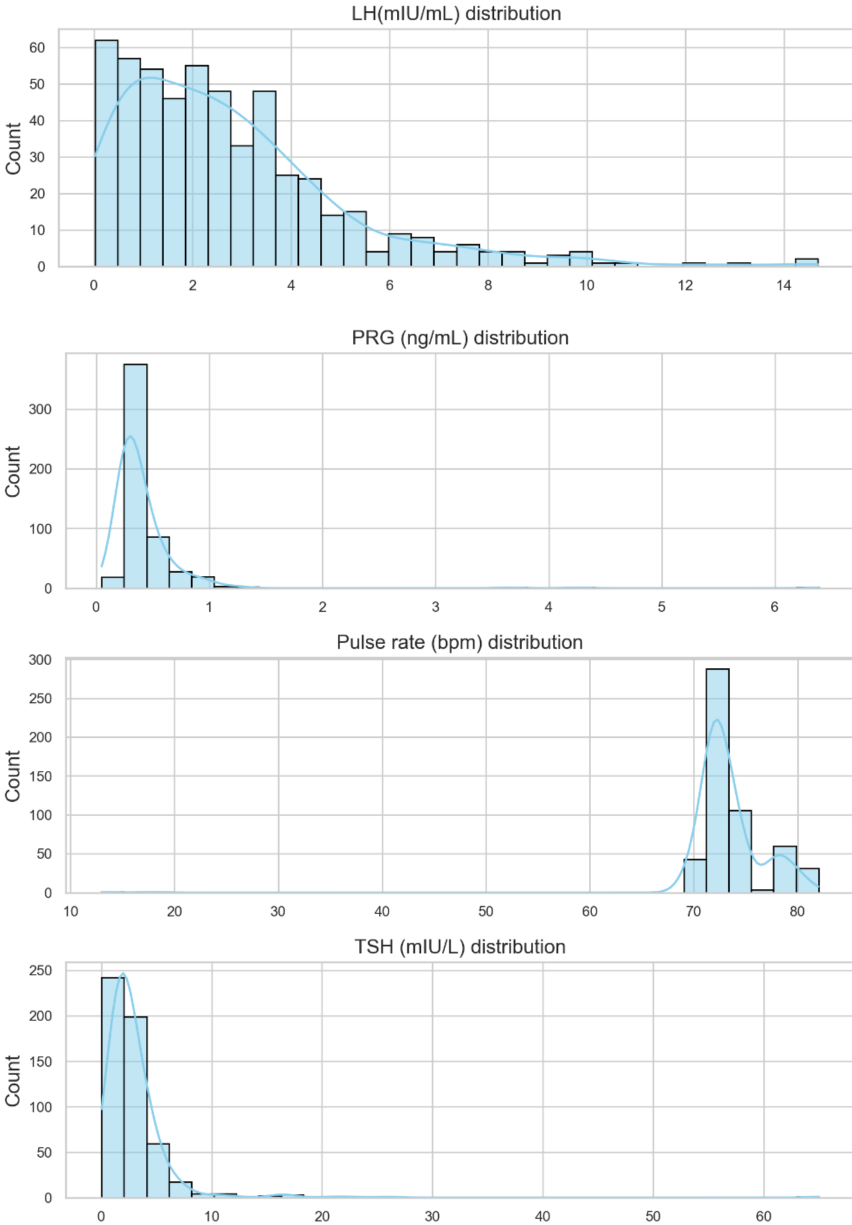


Figure 2d: Distributions of numeric data.

3.2. Train/Test Split

The process of dividing data into training and testing sets is a fundamental concept in machine learning, playing a crucial role in determining a model's performance and its ability to generalize to unseen data. An important factor to consider when splitting the dataset is its impact on the model's generalization capabilities. The choice of the train-test split ratio is also widely debated, with common practices favoring proportions such as 70:30 or 80:20 for training and testing, respectively (Yılmaz, 2024).

K-fold cross-validation is a fundamental technique employed in the evaluation of machine learning models, providing a robust framework for assessing their generalization capabilities. This method involves partitioning the dataset into 'k' subsets, or folds, where the model is trained on 'k-1' folds and validated on the remaining fold. This process is repeated 'k' times, with each fold serving as the validation set once, thereby ensuring that every data point is utilized for both training and validation (Aprihartha & Idham, 2024; Lumumba, Kiprotich, Mpaine, Makena, & Kavita, 2024). The primary advantage of k-fold cross-validation lies in its ability to mitigate the risks of overfitting and underfitting, which are common pitfalls in model training. By averaging the performance across multiple folds, it offers a more reliable estimate of the model's predictive performance on unseen data (Aprihartha & Idham, 2024; Wen et al., 2017).

In this study, the value of k for k-fold cross-validation was set to 5.

3.3. Machine Learning Methods Used

In this study, five different machine learning methods were applied to the dataset, and metrics were calculated. The three methods that provided the highest accuracy rates were selected for implementation using a stacking classifier.

3.3.1. K Nearest Neighbors (KNN) Classifier

The K-Nearest Neighbors (KNN) algorithm, a widely recognized supervised learning technique, falls under instance-based learning methods and is particularly valued for its simplicity and effectiveness in classification tasks. KNN operates by classifying a data point based on the predominant class of its k closest neighbors within the feature space, relying on the assumption that similar instances are spatially proximal. Unlike traditional learning algorithms, KNN does not build an explicit model during training but instead retains the dataset in memory, classifying instances only when queried. This characteristic categorizes KNN as a lazy learner. The algorithm typically employs Euclidean distance as its primary metric, though other measures like Manhattan, Minkowski, or Hamming distances may be applied, depending on the nature of the data.

To classify a new data point, KNN calculates its distance from all training instances, identifies the k nearest neighbors, and assigns a class label through a majority vote among them. It supports both numerical and categorical data and can be adapted for multi-class classification problems. Despite its flexibility, KNN can be computationally expensive for large datasets due to the need for distance computation across all training samples. Additionally, the algorithm is sensitive to noise and irrelevant features, which can affect its performance. Nevertheless, KNN remains a widely used and intuitive classification approach, making it a favorable choice for various applications. A solid understanding of its underlying principles is essential for researchers and practitioners aiming to leverage its potential in practical scenarios (Yılmaz, 2024).

3.3.2. Logistic Regression (LR) Classifier

Logistic Regression (LR) is a widely utilized statistical approach for binary classification, designed to estimate the probability of a binary outcome based on one or more independent variables. Its flexibility in handling non-linear relationships between the dependent and independent variables makes it suitable for applications across various domains, including medicine, finance, and social sciences. A notable challenge in employing LR is addressing class imbalance, where one class is underrepresented in the dataset. This imbalance can lead to biased parameter estimates and reduced predictive performance, as the model may disproportionately favor the majority class. Strategies such as undersampling the majority class or oversampling the minority class are commonly implemented to mitigate this issue.

LR is particularly valued in applications like medical diagnosis, credit risk assessment, and marketing due to its computational efficiency and ability to process large datasets, making it well-suited for real-time decision-making. However, the method is sensitive to outliers, which can skew the model and result in inaccurate predictions. Additionally, its performance relies on the underlying assumptions of the logistic model, which, if violated, may affect reliability. Despite these challenges, LR remains a cornerstone in statistical modeling and machine learning, appreciated for its interpretability and adaptability. To ensure robust and reliable predictions, practitioners should carefully account for potential limitations, including class imbalance and sensitivity to outliers, while applying this technique (Yılmaz, 2024).

3.3.3. Random Forest (RF) Classifier

The Random Forest (RF) classifier is a powerful ensemble learning method designed to enhance classification accuracy and mitigate overfitting by leveraging multiple decision trees. First introduced by Leo Breiman in 2001, RF has become widely used for its robustness and adaptability in various domains, including ecology, finance, and healthcare. The algorithm constructs a collection of decision trees during the training phase, each derived from a random subset of the training data through bootstrap aggregating (bagging). This process reduces variance and improves the model's generalization capability. Additionally, RF introduces randomness by selecting a subset of features at each node split, which fosters diversity among the trees and further reduces the risk of overfitting.

RF offers numerous benefits, such as high classification accuracy and the ability to model complex relationships within the data. By averaging the predictions of individual trees, it minimizes the risk of overfitting compared to standalone decision trees. Furthermore, RF is well-suited for handling missing data and can operate effectively in real-world settings where data quality might be inconsistent. Despite these advantages, training a large number of trees can be computationally intensive, especially when working with high-dimensional datasets. Additionally, the model's performance can be sensitive to hyperparameter configurations, such as the number of trees and the maximum depth of each tree, requiring careful tuning to achieve optimal results. Nonetheless, RF remains a versatile and reliable classifier for diverse machine learning applications (Yılmaz, 2024).

3.3.4. CatBoost (CB) Classifier

The CatBoost (CB) classifier, developed by Yandex, is a highly advanced machine learning algorithm designed for gradient boosting using decision trees. A key distinguishing feature of CatBoost is its ability to natively handle categorical features without requiring extensive preprocessing, unlike other gradient boosting frameworks such as XGBoost and LightGBM. This unique capability enhances its efficiency and ease of use, particularly in datasets where categorical variables are prevalent.

One of CatBoost's most innovative features is its "ordered boosting" technique, which mitigates overfitting and improves generalization. This approach leverages a permutation-based strategy to structure the training process, considering the sequential order of data points to produce more robust models. Additionally, CatBoost employs a specialized encoding mechanism for categorical variables, enabling automatic transformation of these features into a suitable format without manual intervention, streamlining the modeling process.

CatBoost has demonstrated exceptional performance across a wide range of applications, including finance, healthcare, and cybersecurity. For instance, it has been successfully applied in predicting cardiac surgery-associated acute kidney injury, achieving high ROC-AUC scores and proving its effectiveness in medical analytics. Similarly, its utility in detecting botnet attacks underscores its adaptability and reliability in cybersecurity tasks. These examples highlight CatBoost's versatility and its strong performance in complex, real-world scenarios (Yılmaz, 2024).

3.3.5. Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) classifier is a type of artificial neural network structured with multiple layers of neurons, comprising an input layer, one or more hidden layers, and an output layer. MLPs are widely employed in supervised learning tasks, such as classification and regression, owing to their ability to capture complex patterns in data through non-linear mappings facilitated by activation functions.

An MLP consists of interconnected nodes (neurons) arranged in sequential layers. Each neuron receives inputs from the preceding layer, computes a weighted sum, and applies a non-linear activation function—such as sigmoid, hyperbolic tangent (tanh), or Rectified Linear Unit (ReLU)—to produce an output. The neurons in the output layer generate class probabilities for classification tasks. MLP training involves a process called backpropagation, where the weights of the connections between neurons are iteratively adjusted to minimize the error between predicted and actual outputs. This optimization is often performed using algorithms like stochastic gradient descent (SGD) or Adam.

MLPs are versatile due to their ability to approximate any continuous function using non-linear activation functions, making them suitable for solving complex problems. Their architecture is highly customizable, allowing users to adjust the number of hidden layers and neurons to tailor the network to specific tasks. Additionally, MLPs can learn hierarchical representations of data directly from raw inputs, reducing the dependency on manual feature engineering.

However, MLPs also come with certain limitations. They are prone to overfitting, particularly when the model is overly complex relative to the size of the training dataset. Techniques such as dropout, regularization, and early stopping are commonly applied to address this challenge. Moreover, training MLPs can be computationally demanding and time-intensive, especially for large datasets or deep architectures. The performance of MLPs is highly sensitive to hyperparameter choices, including learning rate, number of layers, and neurons per layer, necessitating careful tuning to achieve optimal results. Despite these challenges, MLPs remain a cornerstone of neural network models, offering powerful solutions for a broad range of applications (Yılmaz, 2024).

3.3.6. Stacking Classifier

Stacking classifiers, also known as stacked generalization, is an advanced ensemble learning technique that aims to improve predictive performance by combining multiple base classifiers. The fundamental principle behind stacking is to leverage the strengths of various models to create a more robust and accurate predictive system. In a typical stacking architecture, several base classifiers are trained on the same dataset, and their predictions are subsequently used as input features for a higher-level meta-classifier. This meta-classifier synthesizes the outputs of the base classifiers to produce a final prediction, effectively capturing the diverse perspectives of the individual models (Álvarez, Sierra, Arruti, López-Gil, & Garay-Vitoria, 2015; Kasthuriarachchi & Liyanage, 2021).

One of the key advantages of stacking classifiers is their ability to reduce the risk of overfitting, which can occur when a single model is overly complex or tailored to the training data. By integrating multiple models, stacking can enhance generalization capabilities, leading to improved performance on unseen data (Ledezma, Aler, Sanchis, & Borrajo, 2010). Moreover, stacking is highly flexible, allowing practitioners to combine different types of classifiers, such as decision trees, support vector machines, and neural networks, which can be particularly beneficial in complex prediction tasks (Awang, Makhtar, Udin, & Mansor, 2021). The choice of base classifiers and the meta-classifier is crucial, as the effectiveness of the stacking ensemble largely depends on the diversity and complementary strengths of the models involved (Kusumaniswar, Chinta, & Shareff, 2024).

In this study, five different machine learning methods were applied to the dataset. The accuracy values calculated as a result of these methods are presented in Table 2.

Table 2: Accuracy results of ML methods.

ML method	**Accuracy value (%)**
Logistic Regression	91.68
K-Nearest Neighbors	82.44
Random Forest	90.20
Multilayer Perceptron	89.65
CatBoost	90.76

For the application of the Stacking Classifier method, the top 3 methods (LR, CB, RF) that provided the highest accuracy values were selected. The metrics calculated as a result of the Stacking Classifier are presented in Table 3.

Table 3: Calculated metrics of Stacking Classifier.

Metric	Value (%)
Accuracy	92.52
Sensitivity	82.86
Specificity	97.22
F_1 Score	87.88

Additionally, the results of applying the Stacking Classifier to the dataset, including the confusion matrix, True positive rate (TPR) matrix, and ROC curve, are presented in Figure 3, Figure 4, and Figure 5, respectively.

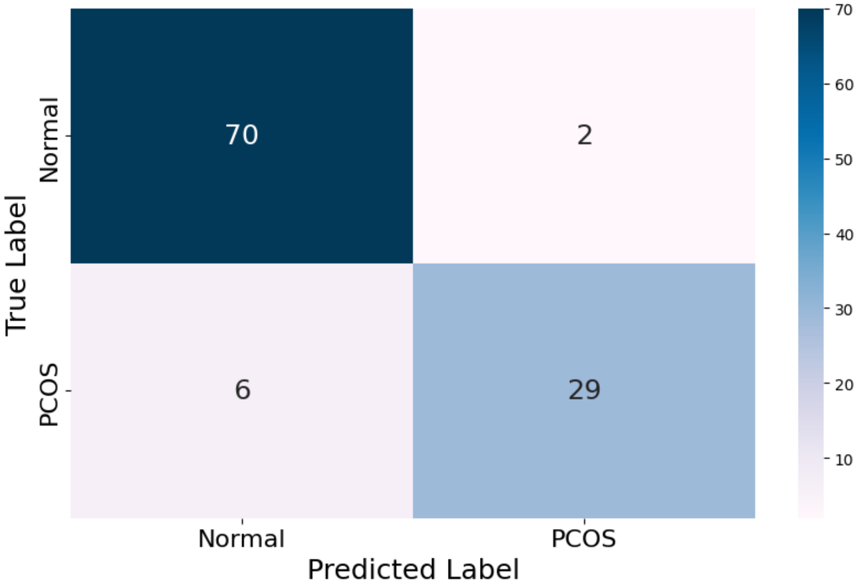


Figure 3: Confusion matrix.

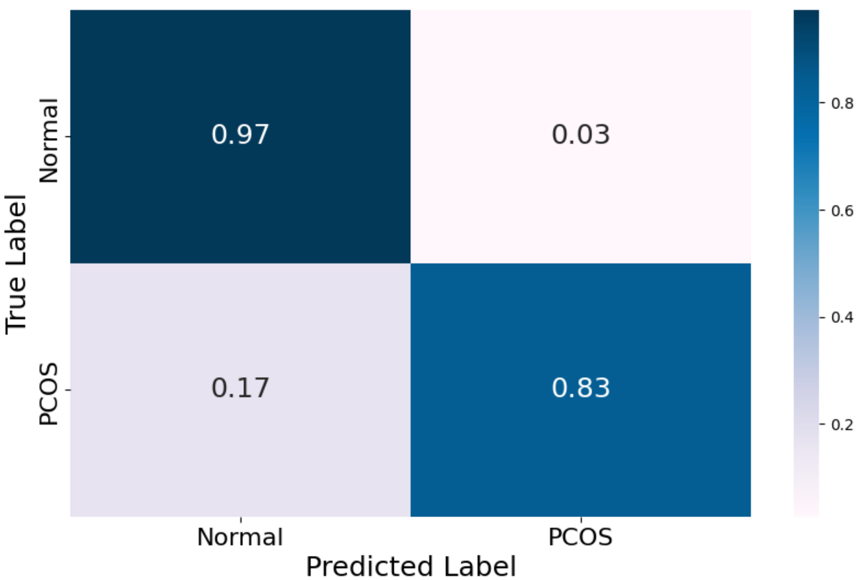


Figure 4: True positive rate matrix.

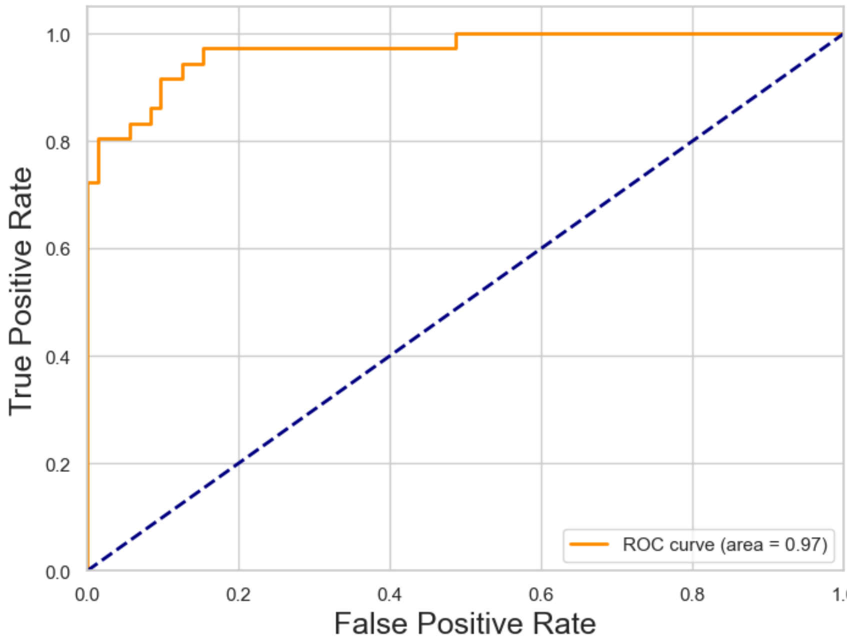


Figure 5: ROC curve and score.

3.3.6. Web-Based Prediction Interface

The most accurate results were obtained using the Stacking Classifier model, which has been preserved to enable predictions. This accomplishment supports the project's objective of creating a web-based prediction system that facilitates user data submission and provides predictive feedback. The implemented web application comprises a backend developed with the Python-based Flask framework, while the frontend is designed using JavaScript. This configuration allows users to input their data and receive predictions regarding PCOS in real time. An illustrative screenshot of the developed interface is provided in Figure 6.

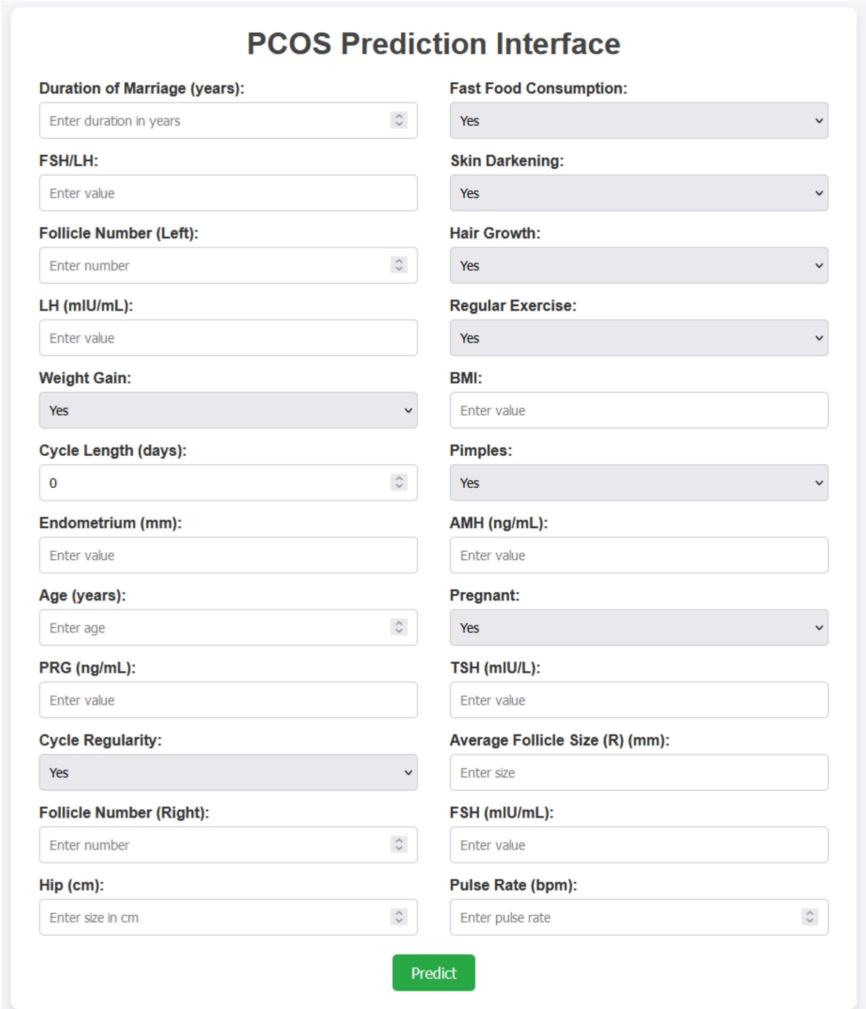


Figure 6: Screenshot of developed web interface.

4. Conclusion

This study presents a methodical and innovative approach to developing a predictive model for Polycystic Ovary Syndrome (PCOS), combining advanced data preprocessing techniques with state-of-the-art machine learning methods. The primary goal was to address challenges associated with data complexity and feature redundancy, ensuring the creation of an accurate and accessible prediction system. Through the application of Recursive Feature Elimination (RFE) and Feature Importance (FI), the original dataset of 42 features was reduced to a more manageable set of 24 critical features. This reduction not only simplified the data collection process but also enhanced the overall model performance by focusing on the most impactful attributes. The preprocessing phase also included outlier removal, improving data consistency and quality, with the final dataset comprising 534 rows and 25 columns (24 input features and 1 output feature).

The study evaluated the performance of five machine learning classifiers—Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF), CatBoost (CB), and Multilayer Perceptron (MLP). Each method was assessed using a 5-fold cross-validation approach, ensuring robust evaluation and reliable generalization capabilities. Among these, LR, RF, and CB emerged as the top-performing models, achieving accuracy rates of 91.68%, 90.20%, and 90.76%, respectively. These models were further integrated into a Stacking Classifier ensemble to leverage their individual strengths. The Stacking Classifier demonstrated superior performance, achieving an accuracy of 92.52%, along with other metrics such as a sensitivity of 82.86%, specificity of 97.22%, and an F1 score of 87.88%. This highlights the effectiveness of ensemble methods in enhancing predictive accuracy and reliability in complex healthcare datasets.

Beyond model development, this research emphasized practical application by deploying the Stacking Classifier in a web-based prediction interface. Built using a Python-based Flask backend and a JavaScript-powered frontend, this interface allows users to input clinical and physical data for real-time PCOS prediction. This system represents a significant step towards bridging the gap between machine learning research and its implementation in healthcare, offering a user-friendly tool for clinicians and individuals alike. The interactive design ensures accessibility while maintaining the model's high accuracy, providing actionable insights in a timely manner.

The findings of this study underscore the potential of machine learning, particularly ensemble techniques, in addressing critical challenges in medical diagnostics. By combining rigorous data preprocessing, advanced model selection, and user-focused application design, this research not only contributes to the existing body of knowledge but also offers a practical solution for PCOS prediction. Future work could focus on expanding the dataset to include diverse populations, incorporating longitudinal data to enhance predictive capabilities, and integrating the web-based tool into clinical workflows for broader impact. Overall, this study exemplifies the transformative role of machine learning in healthcare, fostering innovation and improving outcomes through data-driven decision-making.

5. References

A, P. A., Jain, T., Namratha, T. M., S, V., & Lakshmi, T. G. (2023). Detection of PCOS using Ensemble Models. *International Journal for Research in Applied Science and Engineering Technology*, *11*(5), 420–425. https://doi.org/10.22214/ijraset.2023.51426

Aishwarya, K., Sunitha, M., & Chandy, V. (2023). Unveiling Novel Frontiers in Polycystic Ovary Syndrome Management. *International Journal For Multidisciplinary Research*, *5*(5), 5761. https://doi.org/10.36948/ijfmr.2023.v05i05.5761

Al Wattar, B. H., Fisher, M., Bevington, L., Talaulikar, V., Davies, M., Conway, G., & Yasmin, E. (2021). Clinical Practice Guidelines on the Diagnosis and Management of Polycystic Ovary Syndrome: A Systematic Review and Quality Assessment Study. *The Journal of Clinical Endocrinology & Metabolism*, *106*(8), 2436–2446. https://doi.org/10.1210/clinem/dgab232

Álvarez, A., Sierra, B., Arruti, A., López-Gil, J.-M., & Garay-Vitoria, N. (2015). Classifier Subset Selection for the Stacked Generalization Method Applied to Emotion Recognition in Speech. *Sensors*, *16*(1), 21. https://doi.org/10.3390/s16010021

Aprihartha, Moch. A., & Idham, I. (2024). Optimization of Classification Algorithms Performance with k-Fold Cross Validation. *EIGEN MATHEMATICS JOURNAL*, *7*(2), 61–66. https://doi.org/10.29303/emj.v7i2.212

Awang, M. K., Makhtar, M., Udin, N., & Mansor, N. F. (2021). Improving Customer Churn Classification with Ensemble Stacking Method. *International Journal of Advanced Computer*

Science and Applications, *12*(11). https://doi.org/10.14569/IJACSA.2021.0121132

Aydın, H. L., Dayan, B. S., & Güngör, N. D. (2023). Polycystic ovary syndrome. *Journal of Controversies in Obstetrics & Gynecology and Pediatrics*, *1*(3), 68–73. https://doi.org/10.51271/JCOGP-0015

Emanuel, R. H. K., Roberts, J., Docherty, P. D., Lunt, H., Campbell, R. E., & Möller, K. (2022). A review of the hormones involved in the endocrine dysfunctions of polycystic ovary syndrome and their interactions. *Frontiers in Endocrinology*, *13*, 1017468. https://doi.org/10.3389/fendo.2022.1017468

G. Usha Kiran, D.Yalla Vali, D. Yaswanth Shankar, G. Lokesh, P.Krishna Chaitanya, & S. Sujitha. (2023). Overview Of Polycystic Ovary Syndrome (PCOS). *World Journal of Advanced Engineering Technology and Sciences*, *8*(2), 011–022. https://doi.org/10.30574/wjaets.2023.8.2.0052

Ganie, M., Vasudevan, V., Wani, I., Baba, M., Arif, T., & Rashid, A. (2019). Epidemiology, pathogenesis, genetics & management of polycystic ovary syndrome in India. *Indian Journal of Medical Research*, *150*(4), 333. https://doi.org/10.4103/ijmr.IJMR_1937_17

Ghobrial, S., Ott, J., Steininger, J., Dewailly, D., & Prager, G. (2023). Outcome of Gastric Bypass Surgery on Patients with Polycystic Ovary Syndrome: A Review. *Journal of Clinical Medicine*, *12*(12), 3940. https://doi.org/10.3390/jcm12123940

Glintborg, D., & Andersen, M. (2010). An update on the pathogenesis, inflammation, and metabolism in hirsutism and polycystic ovary syndrome. *Gynecological Endocrinology*, *26*(4), 281–296. https://doi.org/10.3109/09513590903247873

Gupta, A. R., & Agrawal, J. (2021). *Machine Learning-Based Intrusion Detection System With Recursive Feature Elimination*. 157–172. https://doi.org/10.1007/978-981-33-4305-4_13

Gupta, R. K., Kleinjans, J., & Caiment, F. (2021). Identifying Novel Transcript Biomarkers for Hepatocellular Carcinoma (HCC) Using RNA-Seq Datasets and Machine Learning. *BMC Cancer*, *21*(1). https://doi.org/10.1186/s12885-021-08704-9

Hiam, D., Moreno-Asso, A., Teede, H. J., Laven, J. S. E., Stepto, N. K., Moran, L. J., & Gibson-Helm, M. (2019). The Genetics of Polycystic Ovary Syndrome: An Overview of Candidate Gene Systematic Reviews and Genome-Wide Association Studies. *Journal of Clinical Medicine*, *8*(10), 1606. https://doi.org/10.3390/jcm8101606

Jeon, H., & Oh, S. (2020). Hybrid-Recursive Feature Elimination for Efficient Feature Selection. *Applied Sciences*, *10*(9), 3211. https://doi.org/10.3390/app10093211

Kadam, S. (2024). Diagnosis of Polycystic Ovary Syndrome (PCOS) using Deep Learning and Classification Technique's. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, *08*(05), 1–5. https://doi.org/10.55041/IJSREM35318

Kasthuriarachchi, K. T. S., & Liyanage, S. R. (2021). Three-Layer Stacked Generalization Architecture With Simulated Annealing for Optimum Results in Data Mining: *International*

Journal of Artificial Intelligence and Machine Learning, *11*(2), 1–27. https://doi.org/10.4018/IJAIML.20210701.oa10

Khamoshina, M. B., Artemenko, Y. S., Bayramova, A. A., Ryabova, V. A., & Orazov, M. R. (2022). Polycystic ovary syndrome and obesity: A modern paradigm. *RUDN Journal of Medicine*, *26*(4), 382–395. https://doi.org/10.22363/2313-0245-2022-26-4-382-395

Kottarathil, P. (2020). *Polycystic ovary syndrome (PCOS)* [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos

Kumar, N. A. R., & Varadarajan, V. (2024, September 12). *Hybrid PCOS net: A synergistic CNN-LSTM approach for accurate polycysti covary syndrome detection*. https://doi.org/10.21203/rs.3.rs-4311827/v1

Kumari, A., & Jha, S. N. (2024). Unraveling the Complexities of Polycystic Ovarian Syndrome: A Comprehensive Review. *Sumatera Medical Journal*, *7*(2), 96–105. https://doi.org/10.32734/sumej.v7i2.16035

Kurniawati, E. Y., Pramono, N., Hidayat, S. T., & Mahati, E. (2024). Experimental animal models for polycystic ovarian syndrome (methods, effects, and implications). *Livestock and Animal Research*, *22*(1), 67. https://doi.org/10.20961/lar.v22i1.79197

Kusnawi, K. (2024). Leveraging Various Feature Selection Methods for Churn Prediction Using Various Machine Learning Algorithms. *Joiv International Journal on Informatics Visualization*, *8*(2), 897. https://doi.org/10.62527/joiv.8.2.2453

Kusumaniswar, V., Chinta, M., & Shareff, Sk. N. (2024). Bug Classification Using Stacking Classifier. *2024 2nd International Conference on Disruptive Technologies (ICDT)*, 1259–1265. Greater Noida, India: IEEE. https://doi.org/10.1109/ICDT61202.2024.10489044

Ledezma, A., Aler, R., Sanchis, A., & Borrajo, D. (2010). GA-stacking: Evolutionary stacked generalization. *Intelligent Data Analysis*, *14*(1), 89–119. https://doi.org/10.3233/IDA-2010-0410

Liu, N., & Zhu, H. (2024). Progress of Chinese and Western Medicine Research on Polycystic Ovary Syndrome. *Journal of Contemporary Medical Practice*, *6*(6), 79–85. https://doi.org/10.53469/jcmp.2024.06(06).14

Lumumba, V., Kiprotich, D., Mpaine, M., Makena, N., & Kavita, M. (2024). Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models. *American Journal of Theoretical and Applied Statistics*, *13*(5), 127–137. https://doi.org/10.11648/j.ajtas.20241305.13

Misso, M. L., & Teede, H. J. (2015). Metformin in women with PCOS, CONS. *Endocrine*, *48*(2), 428–433. https://doi.org/10.1007/s12020-014-0394-8

Misso, M., Tassone, E., Costello, M., Dokras, A., Laven, J., Moran, L., … on behalf of the International PCOS Network. (2018). Large-Scale Evidence-Based Guideline Development Engaging the International PCOS Community. *Seminars in Reproductive Medicine*, *36*(01), 028–034. https://doi.org/10.1055/s-0038-1667312

Napa, K. K. (2024). Early Prediction of Chronic Heart Disease With Recursive Feature Elimination and Supervised Learning Techniques. *Iaes International Journal of Artificial Intelligence (Ij-Ai)*, *13*(1), 730. https://doi.org/10.11591/ijai.v13.i1.pp730-736

Nethra Sai, M., Sakthivel, V., Prakash, P., Vishnukumar, K., & Dugki, M. (2024). Advanced Diagnosis of Polycystic Ovarian Syndrome using Machine Learning and Multimodal Data Integration. *International Journal of Advanced Computer Science and Applications*, *15*(6). https://doi.org/10.14569/IJACSA.2024.01506122

Peña, A. S., Teede, H., Hewawasam, E., Hull, M. L., & Gibson-Helm, M. (2022). Diagnosis experiences of adolescents with polycystic ovary syndrome: Cross-sectional study. *Clinical Endocrinology*, *96*(1), 62–69. https://doi.org/10.1111/cen.14604

Priyatno, A. M. (2023). Impurity-Based Important Features for Feature Selection in Recursive Feature Elimination for Stock Price Forecasting. *Jurnal Teknik Industri Terintegrasi*, *6*(4), 1182–1194. https://doi.org/10.31004/jutin.v6i4.17726

Priyatno, A. M. (2024a). A Systematic Literature Review: Recursive Feature Elimination Algorithms. *Jitk (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *9*(2), 196–207. https://doi.org/10.33480/jitk.v9i2.5015

Priyatno, A. M. (2024b). Feature Selection Using Non-Parametric Correlations and Important Features on Recursive Feature Elimination for Stock Price Prediction. *International Journal of Electrical and Computer Engineering (Ijece)*, *14*(2), 1906. https://doi.org/10.11591/ijece.v14i2.pp1906-1915

Sowmiya, S., Umapathy, S., Alhajlah, O., Almutairi, F., Aslam, S., & Ahalya, R. K. (2024). F-Net: Follicles Net an efficient tool for the diagnosis of polycystic ovarian syndrome using deep learning techniques. *PLOS ONE*, *19*(8), e0307571. https://doi.org/10.1371/journal.pone.0307571

Sunil, D., Swathi, P., Chidvi Chaturya, K., Sowjanya, N., & Gowreeswari, K. (2024). Deep Learning Based Polycystic Ovary Syndrome (PCOS) Identification. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, *08*(06), 1–5. https://doi.org/10.55041/IJSREM35655

Thangapriya. (2024). Advanced Feature Selection for Human Physiological State Prediction Using Erfe. *Shodhkosh Journal of Visual and Performing Arts*, *5*(3). https://doi.org/10.29121/shodhkosh.v5.i3.2024.2755

Trent, M., & Gordon, C. M. (2020). Diagnosis and Management of Polycystic Ovary Syndrome in Adolescents. *Pediatrics*, *145*(Supplement_2), S210–S218. https://doi.org/10.1542/peds.2019-2056J

Varanasi, L. C., Subasinghe, A., Jayasinghe, Y. L., Callegari, E. T., Garland, S. M., Gorelik, A., & Wark, J. D. (2018). Polycystic ovarian syndrome: Prevalence and impact on the wellbeing of Australian women aged 16–29 years. *Australian and New Zealand Journal of Obstetrics and Gynaecology*, *58*(2), 222–233. https://doi.org/10.1111/ajo.12730

Wen, Z., Li, B., Kotagiri, R., Chen, J., Chen, Y., & Zhang, R. (2017). Improving Efficiency of SVM k-Fold Cross-Validation by Alpha Seeding. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.10785

Yılmaz, H. (2024, September). *AI-Powered Healthcare Innovations: Rehabilitation, Education, And Early Diagnosis*. Serüven Yayınevi. https://doi.org/10.5281/ZENODO.13885904

Zad, Z., Jiang, V. S., Wolf, A. T., Wang, T., Cheng, J. J., Paschalidis, I. Ch., & Mahalingaiah, S. (2023, August 2). *Predicting polycystic ovary syndrome (PCOS) with machine learning algorithms from electronic health records*. https://doi.org/10.1101/2023.07.27.23293255

Zad, Z., Jiang, V. S., Wolf, A. T., Wang, T., Cheng, J. J., Paschalidis, I. Ch., & Mahalingaiah, S. (2024). Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records. *Frontiers in Endocrinology*, *15*, 1298628. https://doi.org/10.3389/fendo.2024.1298628

Zhao, H., Zhang, J., Cheng, X., Nie, X., & He, B. (2023). Insulin resistance in polycystic ovary syndrome across various tissues: An updated review of pathogenesis, evaluation, and treatment. *Journal of Ovarian Research*, *16*(1), 9. https://doi.org/10.1186/s13048-022-01091-0

CHAPTER V

Cryptography and Application Areas

Timuçin KÖROĞLU[1]

1. Introduction

Cryptography, in its simplest form, aims to ensure the confidentiality of messages or data. Thus, it ensures the security of the transmitted data while guaranteeing that the recipient can access the data from secure sources. In today's world, where data communication takes place among billions of people, individuals must use cryptographic technology to protect their sensitive data against attackers.

The history of cryptography dates back to 2000 BCE. The hidden hieroglyphs of Ancient Egypt and Caesar Cipher in Rome prove that cryptography has been used since ancient times (Sharma et al., 2022).

In the early days of cryptography, the focus was on the design of encryption algorithms and the analysis of how well these designs could ensure security. The significant growth in communication directed cryptographic research towards finding methods to enhance

[1] Lecturer Doctor,Pamukkale University, Computer Programming Department, Denizli/Turkey, Orcid:0000-0002-0674-8277, tkoroglu@pau.edu.tr

encryption security and reduce the need for physically sharing encryption keys. Cryptography, in terms of security, can be divided into two main models: information-theoretic security and computational security.

Information-theoretic security explains that a cryptographic structure cannot be broken even if attackers have unlimited computational power. This is because such structures do not rely on difficulty assumptions, such as discrete logarithm computation. On the other hand, the computational security model is based on difficulty assumptions. In these models, the structure can be broken if an appropriate algorithm is found (Meraouche et al., 2021).

This study aims to provide the reader with a broad perspective on the application areas of cryptographic science and the types of attacks that threaten data privacy. In this context, the basics of cryptography and commonly used algorithms will first be briefly explained. Subsequently, the main subject of the study, the application areas of cryptography and types of attacks, will be elaborated.

2. Fundamentals of cryptography

2.1. Cryptography's main objectives

The common goals presented in studies conducted in the field of cryptographic science can be generalized as follows.

Confidentiality: It is the main goal of cryptography. It ensures that only authorized person(s) can access the data by decrypting the encrypted data using the key.

Authentication: It is the process of granting access to a resource that cannot be accessed without authorization, only by verifying the identity of the authorized person(s).

Data Integrity: It explains enabling authorized persons or groups to modify the data while preventing unauthorized persons from altering the data integrity.

Non-Repudiation: It is the situation where the sender and receiver cannot deny having performed the communication or message exchange themselves.

Access Control: It determines to what extent authorized persons or groups can use the resource they have accessed (Abood & Guirguis, 2018).

2.2. Encryption categories

In traditional cryptography, encryption is divided into two main categories. These are symmetric and asymmetric encryption.

Symmetric encryption is an old method in which the key used at the sender side to encrypt the original text is also used at the receiver side to decrypt the encrypted message. The key must be shared in advance between the sender and receiver via a secure communication channel. However, the possibility of the communication channel being intercepted by attackers during the sharing process can create a security vulnerability. Common symmetric encryption algorithms include DES, Blowfish, RC4, RC5 and AES. Asymmetric algorithms, on the other hand, use two keys: public and private. The public key can be known by anyone without any problems. In contrast, the private key is known only to the recipient. The sender encrypts the original text using the public key and the receiver converts the encrypted text back to the original text using its private key. Common asymmetric encryption algorithms are RSA, ECC and DSA algorithms (Al-Shabi, 2019).

In recent years, encryption algorithms that fall outside the classification of symmetric and asymmetric encryption have been discovered. These are quantum cryptography and homomorphic cryptography algorithms.

The ability of quantum computers to solve complex problems in polynomial time threatens the security of data whose confidentiality, integrity, and authenticity are ensured by traditional cryptographic methods. This situation has led researchers to develop post-quantum algorithms to counter quantum attacks. Research has focused on the implementation and optimization of public key encryption (PKE), key encapsulation mechanisms (KEM), and digital signature algorithms on different devices. The interest and intensity of research on Post-Quantum Cryptography (PQC) in recent years indicate that PQC algorithms will become one of the most significant components of cryptology in the future (Dam et al., 2023).

Homomorphic algorithms enable processing encrypted data without decryption in areas where privacy is critical, such as cloud computing and machine learning. Homomorphic algorithms differ in terms of the level of operations they allow and the types of operations they support. Some homomorphic algorithms allow only a limited number of operations of a single type. Among these, RSA and ElGamal are prominent examples. Additionally, studies have developed homomorphic algorithms capable of performing an unlimited number of additions or evaluating any type of operation (Fully Homomorphic Encryption) (Marcolla et al., 2022).

2.3. Hash functions and their properties

Hash functions transform data of arbitrary length into fixed-length data through mathematical computations. Hash functions always produce the same hash value (message digest) for the same input messages and generate different hash values for different input messages. The original message cannot be reconstructed from the message digest of a hash function. Due to this property, hash functions are considered one-way functions. Hash functions must have the property of not producing the same message digest for two different messages. If a hash function behaves this way, it indicates the existence of a security issue called a collision. One of the important use cases of hash functions is to evaluate data integrity and accuracy. A change in the message digest indicates that the input message has been altered (Pittalia, 2019).

3. Cryptographic applications

3.1. IoT security

IoT devices provide great convenience due to the advantages they offer to users. However, due to the low configuration hardware of IoT devices, they face a series of challenges in security areas such as privacy, authentication in access, data management, etc. For these reasons, one of the biggest issues of IoT devices is security. AES, RSA, ECC, and similar algorithms are not practical for IoT devices due to their high resource consumption. Therefore, security in IoT devices should be ensured with lightweight cryptographic algorithms. These algorithms are more effective in providing security for devices like IoT, which have limited processing power and resources. Lightweight cryptographic algorithms must meet the following standards:

- The key size should be at least 80 bits. However, due to the need for longer-term security in IoT devices with low power consumption, it is recommended that the key length be 112 bits.
- Due to the limited resources of these devices, the code length of the cryptographic algorithm and the amount of required RAM must be less than the recommended resource usage limits.
- The chip area and energy consumption of the devices must be below the current ISO standards (Mousavi et al., 2021).

3.2. Block chains

Blockchain technology, which originated with Bitcoin in 2008, has undergone three generations of development in less than a decade. In its inaugural iteration, blockchain introduced a novel monetary system. This digital currency, referred to as cryptocurrency, initially found application in financial contexts involving cash transactions.

In the second generation, blockchains were employed in a broader range of processes, including those related to stocks, loans, and smart property, which are collectively known as smart contracts.

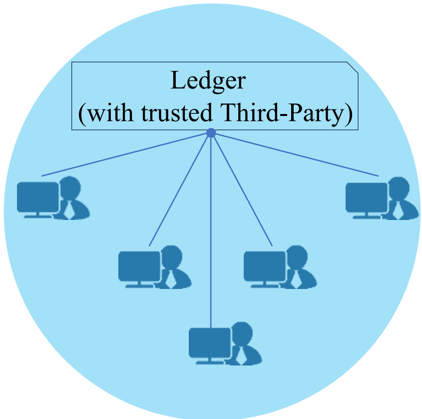


Figure 12. Centralised management with secure third party

Source: (Vaigandla et al., 2023)

In the third generation, blockchain technology began to be utilized in diverse domains such as science, government, and healthcare services.

With the growing prominence of artificial intelligence in recent years, the path has been paved for the fourth generation of blockchain applications (Tripathi, Ahad & Casalino, 2023).

In Figure 1, a representation of traditional ledger technologies requiring a trusted third party is provided. In Figure 2, a representation of the blockchain technology, which does not require the management of a third party, operating on a Peer-to-Peer (P2P) network is provided (Vaigandla et al., 2023).

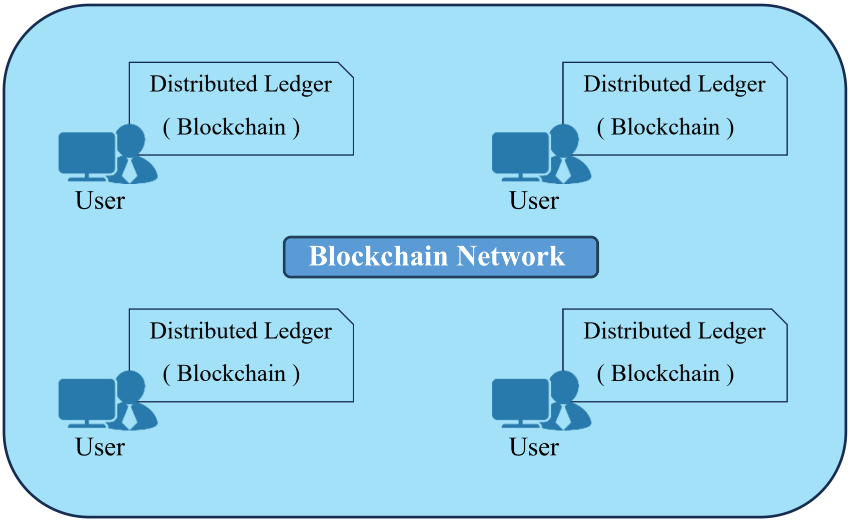


Figure 13. Blockchain's distributed ledger technology

Source: (Vaigandla et al., 2023)

Blockchain is a very large secure structure consisting of blocks that are connected to each other through a specific algorithm, in a continuously expanding manner in a distributed space that is not dependent on a central authority.

Figure 3 shows the general structure of the blockchain. According to this structure, the blockchain structure is divided into three main categories.

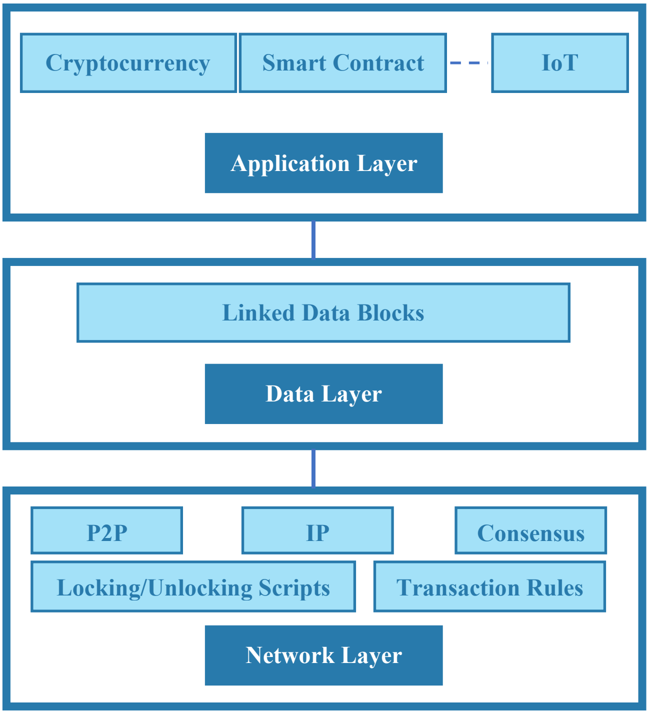


Figure 14. Blockchain general construction

Source: (Gao, Hatcher & Yu, 2018)

3.2.1. Data layer

Consists of many components. The data structures and algorithmic design of these components provide decentralization, continuity, and transparency, which are the most important features of blockchains. According to this design, each block contains a series of information and the hash value that represents the previous block. The inclusion of the previous block's hash value in each block ensures that the blocks are virtually linked to each other. A timestamp containing the creation time is also added to the data stored in the block. Blocks containing a timestamp and information about the previous block become very difficult to alter. Any small change in this information would affect the other blocks in the chain, making it easy to detect and locate the change. The data layer consists of the following components.

3.2.1.1 Data records (transactions)

A large portion of the blockchain data resides in this layer. All transactions, along with their timestamps, are stored in this layer. These transactions include fund transfers or the recording of data to a database, etc.

3.2.1.2 Hash and hash pointer

Blocks are represented by a summary data, where the hash summaries of the important data they contain are combined. Each block stores the summary data of the previous block's hash through a hash pointer. This ensures the logical connection between blocks.

3.2.1.3 Digital signatures

The authentication and integrity of the transactions in the data layer are performed via digital signatures using asymmetric cryptosystem methods.

3.2.1.4 Merkle tree

Transactions in the blockchain are stored in a tree structure. According to this structure, each parent hash is connected to its leaves. Thus, any change in the leaves would affect the root hash value, making the root value used as a defining identifier. A change in the root value indicates the disruption of data integrity.

3.2.1.5 Data blocks

Composed of the block header and the block body. The block header consists of the summary of the previous block, the Merkle tree root hash value, the timestamp containing time information, and the nonce value used for consensus purposes. The block body contains all the transactions.

3.2.2. Network layer

This layer explains the P2P network structure and the consensus structure.

In a P2P network, all users (computers) are equal. These users are called nodes. However, some of these nodes may have higher-level roles. One of these is the "Full Node" role. Such a computer takes the responsibility of holding a copy of the entire blockchain. Therefore, with a single "Full Node," the entire blockchain is protected.

The consensus mechanism builds mechanisms to facilitate consensus for validating each new block added to the blockchain. One of these is Proof of Work (PoW). According to this mechanism, miners calculate a block’s hash according to a very complex chain of rules to add a valid block to the blockchain. This process continues by trial and error until the rules chain is obtained. This operation requires significant time and energy costs.

3.2.3. Application Layer

This layer explains that the blockchain framework can be adapted to other applications beyond cryptocurrency applications like Bitcoin. These include applications such as the Internet of Things (IoT), cloud computing, supply chain tracking, smart contracts, and medical computing (Gao, Hatcher & Yu, 2018).

3.3. Digital Signature

Digital signatures are frequently used in areas such as email transmissions and financial transactions conducted in digital environments (Deng et al., 2022).

Written signatures prove that the person transmitting a document or a data source is the actual sender. Digital signatures, on the other hand, are the digital counterparts of written signatures that fulfill the same function in digital environments. Additionally, digital signatures ensure the integrity of the signed message.

The digital signature creation process consists of two stages: signature generation and signature verification. A digital signature requires the presence of a private key and a public key. There is a mathematical relationship between the private key and the public key. These keys are the responsibility of the signer. In a digital signature system, the private key is not known by anyone other than the signer. The public key, however, can be known by others but its integrity must be preserved.

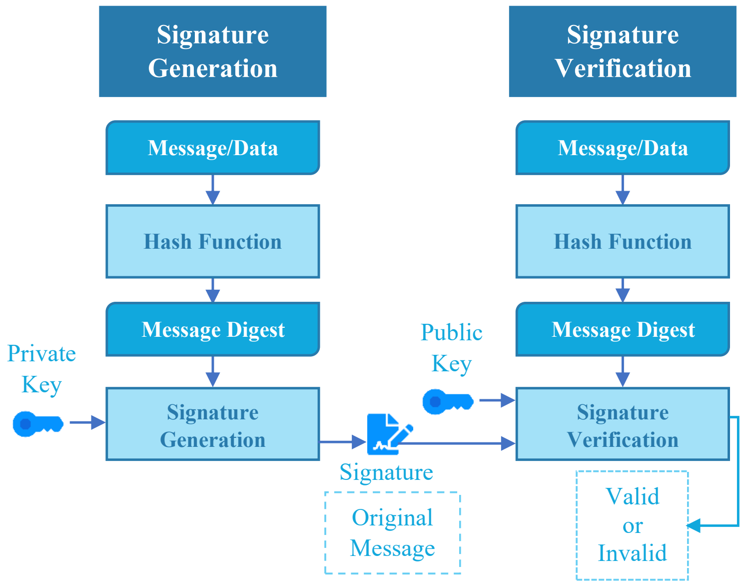


Figure 15. Digital signature workflow

Source: (Chen et al., 2023)

As shown in Figure 4, the private key is only used during the signature generation phase. In this phase, the signer obtains the hash of the message to be transmitted using a hash function and then encrypts the message hash with the private key. This process is referred to as signing. The encrypted message hash with the private key and the original message are then sent to the recipient.

The recipient uses the public key to decrypt and verify the hash. Verifying the hash with the public key means verifying the signer's identity because the private key and the public key are mathematically related, and this relationship is unique. The recipient's next step is to rehash the original message using the same hash function and compare the obtained hash with the hash sent by the signer. If the comparison is successful, the integrity of the data is proven.

The recipient must ensure that the public key truly belongs to the claimed signer, and this assurance is provided through the

Public Key Infrastructure (PKI) mechanism. This function is performed by certificates. Certificates provide a structure where user identity information, the user's public key, the certificate validity period, and similar data are digitally signed using asymmetric encryption algorithms. The purpose is to secure the identity of the public key owner used in the digital signature, thereby preventing fraud and potential security vulnerabilities.

Once the certificate ensures that the public key truly belongs to the signer, if the digital signature can be opened with the public key, it proves that the digital signature is valid and belongs to the rightful owner.

The reasons why the described assurances are needed can be summarized as follows:

If there is no clear assurance that the public key used for signature verification belongs to the signer, attackers with mathematically valid key pairs can forge digital signatures under a false identity and misrepresent their identity for gain.

If, for any reason, the public key loses its mathematical validity, individuals other than the signer may be able to generate signatures that can be verified with that public key.

If the PKI fails to reliably identify the owner of the public key, the public key can be associated with a false identity and used by attackers. Attackers can also associate the signer's identity with a public key that does not belong to the signer. Both cases create security vulnerabilities (Chen et al., 2023).

3.4. Cloud Computing

Cloud computing is a secure structure that provides computing services such as software, enterprise networking, storage services, databases, web, or software server services, etc., to its users at low costs through web services. Its biggest advantage is that, unlike local machines, it allows access to the computing services obtained from any location. Security is a very important issue in cloud computing, which provides all these advantages. The management of the data stored in the cloud structure is not directly carried out by the service recipients. Users do not know if their data is accessed by third parties or what security measures are taken to protect their data. The security of the services provided in cloud computing is ensured through cryptographic algorithms (Bhargav & Manhar, 2020).

Cloud computing provides many advantages to its users. However, cloud computing systems can lead to security vulnerabilities for certain reasons. These reasons are listed below:

Insecure Interface: Users who receive services from cloud computing use cloud computing interfaces and applications to manage many services such as identity management, access management, and data arrangement. Interfaces that are configured insecurely can lead to significant security vulnerabilities in cloud computing systems.

Malicious Cloud Computing Employees: Cloud computing service providers have many employees within their organizations. Security vulnerabilities may occur due to these employees. Users who transfer their data from their local devices to the cloud computing system entrust the security of all their data to the cloud computing system, and consequently, to the employees of the organization. Organization employees may misuse user data or share it with other organizations. Users are unaware of this situation, and significant security problems may arise.

Data Loss and Leakage: The data of users who receive services from cloud computing undergoes two significant changes. The first is the transfer of user data from their own devices to the cloud computing system. The second is the transition from single application mode to multi-application mode. Using data in multi-application mode requires the sharing of data among multiple applications, services, or processes. In this case, data isolation may not be achieved, leading to security vulnerabilities.

Flood Attacks: These types of attacks occur when malicious individuals send fake requests that excessively occupy the cloud computing system. The attackers aim to overload the system, slowing down the service provided to legitimate users.

DDoS Attacks: The purpose of DDoS attacks is also to disrupt the system's operation by overloading the server with a large number of fake requests, similar to flood attacks. As a result, just like in flood attacks, the service received by users either slows down or completely stops. DDoS attacks are more comprehensive than flood attacks. DDoS attacks are carried out from many different devices worldwide, whereas flood attacks are conducted with a more limited number of devices. Flood attacks are a specific type of DDoS attack.

IP Spoofing: The main goal of this attack is to analyze network traffic and enable the attacker to obtain data suitable for their purpose. For this purpose, the attacker identifies the IP address of a trusted device, modifies the packet data, and sends it to the server. The server sees the IP address of the incoming packets as reliable. In this case, communication between the attacker and the server develops in the direction desired by the attacker, allowing them to obtain sensitive data and exploit a security vulnerability (Gupta et al., 2021).

In cloud computing systems, user data is not transferred to the system in its raw form. User data must be encrypted using cryptographic algorithms during processes such as transmission, storage, and data processing in the cloud system (Can et al., 2023).

Security in cloud computing is classified in Figure 5, and the explanations are provided below.

Privacy Protection: Since users in cloud computing will access cloud resources remotely, ensuring the privacy of user data is extremely important. In this context, privacy protection focuses on issues such as protecting user identity, query privacy, and access design security.

Storage Security: This security concept explains the secure storage of user data in the cloud computing system. Users have entrusted the security of their data to cloud providers. In this context, cloud providers must also ensure the integrity of user data.

Data Security: In this protection, three situations are considered:

- **Data Integrity:** In this context, data can only be modified by its legal owner. It is ensured that no one other than the data owner can modify the data. Data integrity is ensured through various cryptographic techniques such as hashing and digital signatures.
- **Access Control:** The access control structure guarantees the access of legitimate users to cloud computing resources while preventing attackers from capturing these resources and using them maliciously.
- **Attribute-Based Encryption:** In this context, access is provided based on the attributes assigned to users. Since the data is encrypted according to user attributes, access is granted only to users with the appropriate attributes (Sujithra et al., 2020).

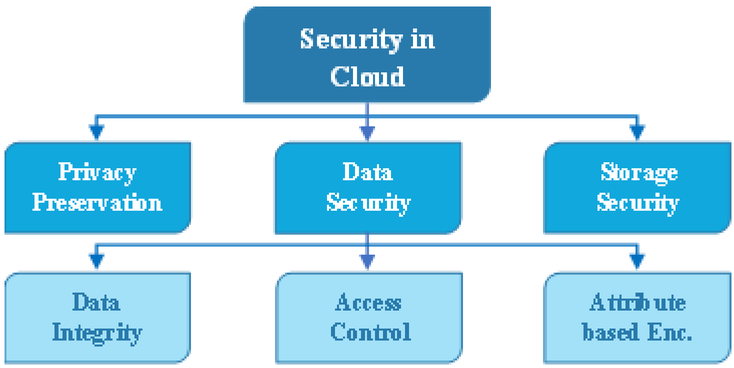


Figure 16. Classification of Cloud Computing Security

Source: (Sujithra et al., 2020)

4. Post-Quantum Cryptography

Today, many widely used cryptographic algorithms are based on the assumption that difficult problems cannot be solved with polynomial complexity. These include factoring a very large integer into its prime factors, solving the discrete logarithm problem, and order-finding. Solving these problems requires exponential time complexity (O(n * 2^n)). It is very difficult for attackers to break encryption within the time output of an exponential function. However, quantum computers, with their quantum-based working principles, can solve these problems with polynomial time complexity of O(n^2).

It should also be taken into account that systems using symmetric encryption algorithms rely on public-key algorithms to securely deliver their secret keys to the recipient. This situation can also lead to vulnerabilities in symmetric encryption algorithms.

Shor's algorithm can solve the factorization of an integer with O(n^2) complexity. Order-finding and discrete logarithm problems have also been solved by other researchers with the same complexity. All these developments indicate that cryptographic algorithms will undergo a significant transformation in the post-quantum computer era (Kumar & Pattnaik, 2020).

5. Conclusion

Cryptography is a subfield of cryptology. Its main function is to transform the original data into a very different content using mathematical operations and to prevent malicious individuals from accessing the original data content. This explains data confidentiality and holds a very important place in the field of security. In addition to ensuring data confidentiality, cryptology is also used for purposes such as data integrity, authorization, access control, and non-repudiation.

The objectives of cryptography in data security have materialized in areas such as the Internet of Things (IoT), blockchain applications, digital signatures, and cloud computing. In these fields, numerous cryptographic algorithms have been developed from past to present to ensure data security. These algorithms, which can be described as traditional, are sufficient in terms of security in today's world. However, the widespread adoption of quantum computers in the near future is anticipated. Due to the high computational power of quantum computers, it is inevitable that the data security provided by traditional cryptographic algorithms will be compromised. Post-quantum cryptography (PQC) studies conducted in recent years aim to eliminate this threat and ensure post-quantum security.

Cryptography has found many different application areas. In this book chapter, cryptography's most widely used areas, such as digital signature applications, cryptographic applications in the Internet of Things, blockchain technology applications, and cryptography in cloud computing, have been presented to the reader with a straightforward explanation.

REFERENCES

Sharma, D. K., Singh, N. C., Noola, D. A., Doss, A. N., & Sivakumar, J. (2022). A review on various cryptographic techniques & algorithms. *Materials Today: Proceedings*, 51, 104-109.

Meraouche, I., Dutta, S., Tan, H., & Sakurai, K. (2021). Neural networks-based cryptography: A survey. *IEEE Access*, *9*, 124727-124740.

Abood, O. G., & Guirguis, S. K. (2018). A survey on cryptography algorithms. *International Journal of Scientific and Research Publications*, *8*(7), 495-516.

Al-Shabi, M. A. (2019). A survey on symmetric and asymmetric cryptography algorithms in information security. *International Journal of Scientific and Research Publications (IJSRP)*, *9*(3), 576-589.

Dam, D. T., Tran, T. H., Hoang, V. P., Pham, C. K., & Hoang, T. T. (2023). A survey of post-quantum cryptography: Start of a new race. *Cryptography*, *7*(3), 40.

Marcolla, C., Sucasas, V., Manzano, M., Bassoli, R., Fitzek, F. H., & Aaraj, N. (2022). Survey on fully homomorphic encryption, theory, and applications. *Proceedings of the IEEE*, *110*(10), 1572-1609.

Pittalia, P. P. (2019). A comparative study of hash algorithms in cryptography. *International Journal of Computer Science and Mobile Computing*, *8*(6), 147-152.

Mousavi, S. K., Ghaffari, A., Besharat, S., & Afshari, H. (2021). Security of internet of things based on cryptographic algorithms: a survey. *Wireless Networks*, *27*(2), 1515-1555.

Tripathi, G., Ahad, M. A., & Casalino, G. (2023). A comprehensive review of blockchain technology: Underlying principles and historical background with future challenges. *Decision Analytics Journal*, 100344.

Vaigandla, K. K., Karne, R., Siluveru, M., & Kesoju, M. (2023). Review on blockchain technology: architecture, characteristics, benefits, algorithms, challenges and applications. *Mesopotamian Journal of CyberSecurity*, 2023, 73-84.

Gao, W., Hatcher, W. G., & Yu, W. (2018, July). A survey of blockchain: Techniques, applications, and challenges. *In 2018 27th international conference on computer communication and networks (ICCCN)* (pp. 1-11). IEEE.

Deng, X., Zhao, W., Shi, R., Ding, C., & Fan, H. (2022). Quantum Digital Signature with Continuous-Variable. *International Journal of Theoretical Physics*, 61(5), 144.

Chen, L., Moody, D., Regenscheid, A., & Robinson, A. (2023). Digital signature standard (DSS). *Federal Information Processing Standards Publication,* https://www.nist.gov/publications/digital-signature-standard-dss-3.

Bhargav, A. J. S., & Manhar, A. (2020). A review on cryptography in cloud computing. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, 6(6), 225-230.

Gupta, K., Gupta, D., Prasad, S. K., & Johri, P. (2021, March). A Review on Cryptography based Data Security Techniques for the Cloud Computing. *In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 1039-1044). IEEE.

Can, O., Thabit, F., Aljahdali, A. O., Al-Homdy, S., & Alkhzaimi, H. A. (2023). A comprehensive literature of genetics cryptographic algorithms for data security in cloud computing. *Cybernetics and Systems*, 1-35.

Sujithra, T., Sumathi, D. M., Ramakrishnan, M., & Sankar, S. M. (2020). Survey on data security in cloud environment. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, 11(4).

Kumar, M., & Pattnaik, P. (2020, September). Post quantum cryptography (pqc)-an overview. *In 2020 IEEE High Performance Extreme Computing Conference (HPEC)* (pp. 1-9). IEEE.

CHAPTER VI

Motivation and Application of Haptic Systems

ŞAFAK KILIÇ[1]

Research Disciplines

Haptics, in a non-scientific context, pertains to the sense of touch and all aspects associated with it. Upon deeper consideration, it becomes evident that touch inherently requires interaction. As a result, touch perception cannot occur without physical contact, implying that something must either be touched or touch another object. Based on this fundamental principle, it is clear that haptics revolves around interaction. Although this statement seems straightforward, it introduces complexity in research and technical applications. Unlike vision and sound, haptics inherently affects the object being touched due to the nature of interaction. Additionally, the classification of these interactions depends on the physical characteristics of both the body and the object involved. Considering

[1] Dr. Öğr. Üyesi, Kayseri Üniversitesi, Mühendislik Fakültesi, Yazılım Mühendisliği Bölümü, Kayseri/Türkiye, Orcid: 0000-0002-2014-7638, safakkilic@kayseri.edu.tr

that the sense of touch is vital to every mechanical body part interacting with the environment, particularly the skin-covered regions with varying sensory capabilities, the challenges in this domain become evident.

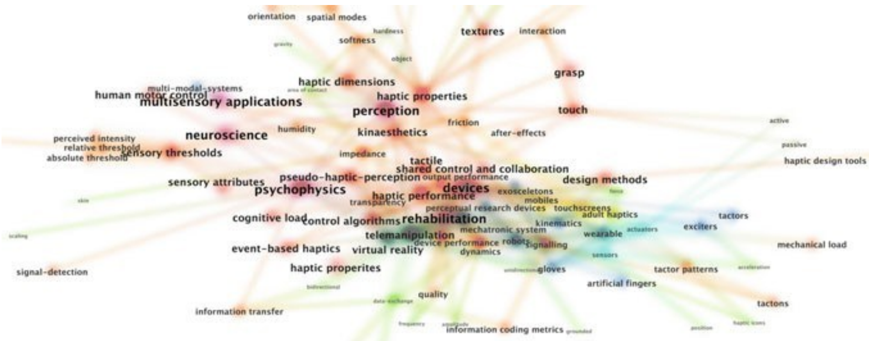


Figure 1.1 Concept-Map on Haptic Disciplines, own visualization

As a continuously evolving field, haptic research undergoes frequent restructuring. Figure 1.1 provides a snapshot of the current core disciplines. While there were perhaps eight to ten areas of haptic research two decades ago, the field has undergone significant diversification in the past ten years. This shift is due to a growing understanding of interdependencies, as well as increased specialization driven by industry demands. One primary focus is perception-based research, which explores psychophysical and neuroscience-related topics. This area significantly influences application-based research, which itself relies on various components and subsystems, each tailored to specific applications.

Some Broad Scope on Haptics

But what exactly is haptics? A widely accepted and general definition describes it as the study of touch and its associated phenomena. This field encompasses not only the sensory experience of touch but also the mechanical and cognitive processes involved in perceiving and interacting with objects through tactile means. It serves as a bridge between physical interaction and perceptual understanding, playing a critical role in both natural and artificial systems. This chapter provides a more detailed exploration of the concept of haptics in Section 1.4, followed by an introduction to four broad categories of applications for haptic systems in Section 1.5. These serve as a foundation for understanding the motivation behind designing haptic systems and, ultimately, for the development of this book. Before delving into these topics, Section 1.3 offers a brief overview of the philosophical and social dimensions of the human sense of touch. While these aspects are not discussed further in the book, they are essential considerations for any engineer working in the field of haptics.

Philosophical and Social Aspects

Engineers often describe haptics using technical terms such as forces, elongations, frequencies, mechanical stresses, and shear forces, which are critical to the design process. However, the scope of haptics extends far beyond these mechanical definitions. Haptic perception spans a broad spectrum, from simple daily interactions—such as drinking from a glass or typing—to social communications like handshakes or a pat on the shoulder, as well as deeply personal and private interpersonal experiences. Touch has both conscious and significant unconscious components. For example, Crusco and Wetzel (1984) demonstrated that a subtle touch could increase a customer's tip to a server by approximately 10%, an effect known as the "Midas Touch." Interestingly, this phenomenon appears to be largely independent of the gender or age of either party. This section examines the wide-reaching influence of haptics on human behavior, beyond its technological definitions. It also urges engineers to approach the design of haptic systems with ethical awareness, recognizing the potential to influence and manipulate human perception through touch.

Haptics as a Physical Being's Boundary

The term "haptics" originates from the Greek word "haptios," meaning "something that can be touched." Over the course of human history, the awareness and understanding of the haptic sense have undergone numerous transformations. Aristotle, for example, ranked touch last among the five primary senses in his philosophical discussions (Wolf, 2007).

sight

hearing

smell

taste

touch

Despite ranking touch last among the five senses, Aristotle acknowledged its fundamental importance as early as 350 B.C., stating that among animals, touch is the most indispensable sense (Adams, 1999). Historically, the social perception of touch has undergone significant shifts. At times, it was stigmatized for its association with physical desires, as reflected in the perspective that sight remains "virgin" in contrast to touch, which transmits lust (Wolf, 2007). In some contexts, touch was even referred to as the "sense of excess" (Grunwald, 2008). Within the general dichotomy of higher and lower senses, touch was often relegated to the latter. In

Western civilization, the Church once deemed touch forbidden due to the pleasures it could provide.

However, public perception began to shift during the 18th century. Kant described touch as the only sense capable of providing direct external perception, considering it the most significant and educational of all the senses, though also the coarsest. He argued that without touch, humans would lack an awareness of their physical form, as this perception underpins the knowledge derived from sight and hearing (Kant, 1983). Kant emphasized the central role of touch in spatial perception, noting that it allows humans to contextualize and understand impressions gathered by other senses. While stereoscopic vision and hearing develop early, the ability to interpret and integrate these perceptions with spatial awareness relies on touch. This unique sense bridges the gap between the self and external objects, with the skin serving as the organ that defines physical boundaries and enables tactile exploration of the environment.

Formation of the Sense of Touch

As discussed in the previous section, the sense of touch serves a variety of purposes. Understanding these functions equips engineers to establish requirements for technical systems. To do so effectively, it is beneficial to consider the diverse roles that the haptic sense fulfills. However, at this stage, rather than measuring its characteristics, the focus is on observing the properties of objects that the sense of touch can differentiate.

Touch is not solely concerned with perceiving the physical boundaries of the body but also plays a crucial role in analyzing the immediate environment, including objects and their attributes. Throughout human evolution, the ability to discriminate between different textures, such as the structure of fruits and leaves, has been essential for survival. For example, humans learned to determine ripeness or edibility by touch, such as distinguishing a furry berry from smooth ones. Similarly, touch enables the recognition of potentially harmful textures, like spiny seeds, allowing careful handling to access the contents without injury.

The sense of touch has evolved to excel at perceiving and differentiating surface properties, including roughness. These properties vary widely, from smooth, ceramic-like or lacquered surfaces with structural dimensions in the micrometer range, to moderately textured surfaces like coated tables, and rough surfaces such as coarse cord textiles with mesh sizes measuring several millimeters.

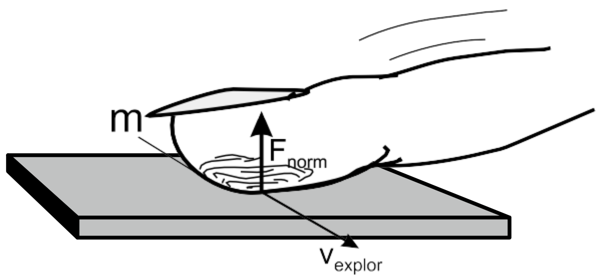


Figure 1.2 Illustration for the interaction of movements, normal forces on the finger pad and frictional coupling

When humans interact with surfaces, they employ a characteristic approach that allows them to gather perceptual information. By sliding their finger across a surface (Fig. 1.2), they create shear forces that act upon the skin. The intensity of these forces is determined by several factors that affect the friction between the skin and surface: the skin's elastic properties in the tangential direction, the normal force (Fnorm) applied during touch, the speed at which the finger moves during exploration (vexplr), and the coefficient of friction (μ) between the surfaces in contact.

Engineers familiar with designing friction-based mechanical couplings understand that achieving a friction coefficient (μr) greater than 0.1 between two surfaces is challenging without incorporating additional structures or adhesive materials. However, nature has developed an ingenious solution to enhance shear force coupling with the skin through fingerprints, particularly on our primary tactile exploration tool. These epidermal ridges effectively transfer shear forces to the skin by creating bending moments in its upper layers through their bar-like structures. Furthermore, these ridges enable mechanical interlocking with surface features of comparable dimensions, essentially creating a form-fitting connection between the skin and the object being handled. While this structural function might seem unexpected initially, it serves as another reminder that nature's designs always serve a specific purpose.

This understanding of shear force coupling to the skin has led to two practical advancements. First, recent research has emphasized the significance of this coupling mechanism, leading to improvements in the design processes of tactile devices (Gerling & Thomas, 2005). Second, this knowledge has been applied to enhance the measurement precision of commercial force sensors by incorporating ridge-like structures (Vasarhelyi et al., 2006).

Another important aspect of the haptic sense, and likely an evolutionary advantage, is its capacity to facilitate the use of tools. Mechanoreceptors in the skin, as described in Section 2.1, are capable of detecting high-frequency vibrations generated during interactions with stiff tools. This ability allows humans to discern surface properties, detect contact points, and identify collisions (Fiene et al., 2006).

Touchable Art and Haptic Aesthetics

In the 20th century, art increasingly engaged with the sense of touch, exploring and challenging its meaning. One striking example is the "Furry Cup" (Fig. 1.3), which draws attention to the importance of haptic texture in perceiving surfaces and surface structures. While the cup's general shape remains visually recognizable, its originally smooth ceramic surface is entirely covered with fur, creating a tactile juxtaposition.

In 1968, the "Pad- and Touch-Cinema" (Fig. 1.4) by Valie Export allowed participants to touch the artist's bare skin for 12 seconds through a curtain-covered box. According to Export, this was the only authentic way to experience sexuality without incorporating voyeurism (Getzinger, 2006). These instances are just a few examples of how art and artists have explored the diverse facets of haptic perception.

Similarly, haptic interaction within virtual environments also carries artistic characteristics. For instance, in 2004, Ishii from the MIT Media Laboratory and Iwata from the University of Tsukuba showcased innovative "tangible user interfaces," where bottles could be opened to "release" music, blending artistry with technology.

Today, devices not only respond to human touch but also "touch back." For example, Marc Teyssier has been actively investigating the boundaries of social acceptability in the intersection of art and robotics (Fig. 1.5). Beyond their artistic value, these installations inspire ongoing research into new interaction possibilities in Human-Computer Interaction (HCI), leveraging concepts born from this fusion of art and technology.

Figure 1.3 features Meret Oppenheim's Furry Cup (1936), an iconic piece of 20th-century art that emphasizes the importance of texture in haptic perception. The work juxtaposes the familiar shape of a ceramic cup with an unconventional fur-covered surface, challenging traditional associations of materiality and touch (Getzinger, 2006; Néret, 1998). The digital image is credited to The Museum of Modern Art/Scala, Florence, Qc 2022.

Figure 1.4 depicts Valie Export's TAPP und TASTKINO (1968), a provocative art installation that allowed participants to touch the artist's bare skin for 12 seconds through a curtained box. This piece challenged traditional notions of voyeurism and redefined the boundaries of haptic interaction in art. The black-and-white photograph is credited to Valie Export, Bildrecht Wien, 2022, with photography by Werner Schulz. Additional information is available courtesy of Valie Export at link.

Figure 1.5 showcases the MobiLimb project, a novel exploration by Marc Teyssier that introduces a device capable of "touching back." This work delves into the boundaries between human and robotic interaction, questioning what is socially acceptable in this emerging field. The image is credited to Marc Teyssier, Qc 2022, and is used with permission (Teyssier et al., 2018).

In one example, picture frames are utilized as tangible objects to initiate video calls with relatives and friends by placing them on a designated area of a specialized tablecloth (Wilde et al., 2013). Similarly, Disney Research introduced Touché, a capacitive sensing technique that transforms nearly any object into a touch input device, aiming to advance immersive computing by embedding touch capabilities into everyday objects (Sato et al., 2012). Another innovative approach is demonstrated by Playtronica, which enhances touch functionality in ordinary objects by interpreting capacitance into MIDI signals and synthesized music (Fig. 1.6).

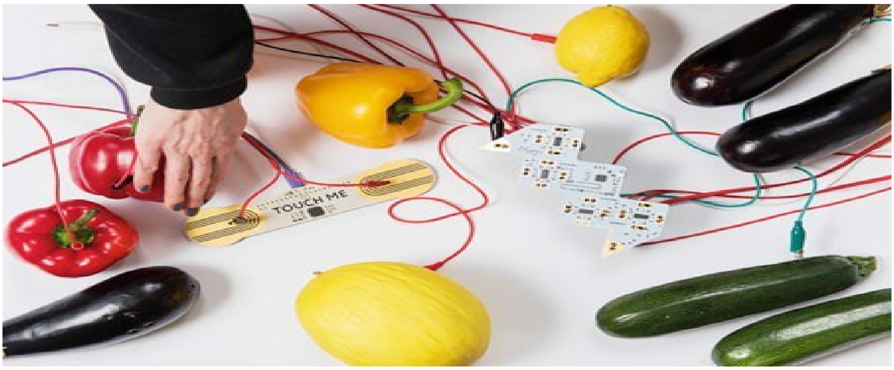


Figure 1.6 showcases Playtronica's Playtron and Touch Me products, which use capacitive measurement to generate MIDI sounds based on touch intensity. Image credited to Daria Malysheva, Qc 2022, used with permission.

In technical applications, haptic aesthetics significantly influence user experience. Automotive manufacturers aim to create a tactile brand identity by developing objective quality standards for interface perception (Anuguelov, 2009; Reisinger, 2009). Some companies claim to "make percepts measurable" (Battenberg Robotic GmbH & KG Co, 2007), while designers provide toolkits to evaluate knobs, switches (Gaspar et al., 2019; Jagodzinski & Wintergerst, 2009), and vibrational feedback through specialized design packages (Israr et al., 2019). However, the mechanisms underlying haptic aesthetics are not yet fully understood. Studies generally employ multidimensional scaling and regression algorithms to link subjective evaluations with objective metrics (Rösler et al., 2009), but the perceptual dimensions and data models remain active research areas (Okamoto et al., 2013; Culbertson & Kuchenbecker, 2017).

Carbon and Jakesch proposed a comprehensive model that connects object properties with familiarity assessments, highlighting the interdisciplinary collaboration of engineering and psychology in product design (Carbon & Jakesch, 2013; Breitschaft & Carbon, 2021).

Technical Definitions of Haptics

Haptic technology relies on a clear understanding of terms and concepts to effectively design and apply sensory systems. Technical definitions and classifications of haptic interactions and perception provide the foundation for such innovations. These frameworks allow for a structured exploration of how humans perceive and interact with haptic systems, encompassing both tactile and kinaesthetic dimensions. By establishing these definitions, researchers and developers create a common language to address the complexities of designing intuitive and functional haptic interfaces. This foundational knowledge sets the stage for more in-depth discussions on perception and interaction mechanisms, as explored in subsequent sections. *Definitions of Haptic Interactions*

The haptic system enables humans to interact with both real and virtual environments by leveraging mechanical, sensory, motor, and cognitive capabilities (Jandura & Srinivasan, 1994). These interactions consist of one or more fundamental operations, which are generally categorized into motion control and perception (Kirkpatrick & Douglas, 2002). These operations are referred to as primitives because they cannot be further subdivided or reclassified.

The perception category includes primitives such as detection, discrimination, identification, and scaling of haptic information (Gall et al., 2001). The study of these primitives falls under the scientific discipline of psychophysics. To further describe the perception primitives, the concept of a stimulus must first be defined.

Definition *Stimulus (pl. stimuli)* Excitation or signal that is used in a psychophysical procedure. It is normally denoted with the symbol Φ. The term is also used in other contexts, when a (haptic) signal without further specification is presented to a user.

Haptic stimuli typically include forces, vibrations, stiffness, or objects with distinct properties. With this understanding, perception primitives can be analyzed more closely, as each primitive is applicable to specific types of haptic stimuli:

- **Detection**: This primitive refers to the process by which a human or user perceives the presence of a stimulus. The ability to detect a stimulus depends on sensory organs (as detailed in Section 2.1) and neural processing. Only detected stimuli can be further processed by other perception primitives.
- **Discrimination**: When multiple stimuli are present, this primitive involves distinguishing differences in their properties, such as vibration frequency or amplitude, or object characteristics like texture, hardness, and mass.
- **Identification**: This primitive builds on detected stimuli by comparing them not with each other but with practical or abstract knowledge. This allows the classification of stimuli,

such as identifying geometric properties like shape or size.

- **Scaling**: Psychophysicists identify scaling as a perception primitive that involves rating the properties of stimuli or objects. While scaling is secondary in interaction analysis, it can provide valuable insights into signal magnitudes during the design process (Stevens, 1975).

Motor control primitives, like perception primitives, are categorized into distinct operations:

- **Travel**: This involves moving limbs, the entire body, or virtual representations (avatars) to explore environments, reach destinations, or reposition oneself. Adjusting ongoing movements is also part of this primitive.
- **Selection**: Particularly in virtual environments, this primitive refers to marking or selecting an object or function, enabling direct interaction with the environment.
- **Modification**: Based on the selection of an object or function, this primitive describes changes in orientation, position, or properties of an object, as well as combining multiple objects into one.

When applying motor control primitives, it is essential to consider both the operation and its objective. For instance, using a mouse to select an icon on a computer screen involves the *travel* primitive to navigate to the object, followed by the *selection* primitive to mark it. If a novel haptic device replaces the mouse, the *travel* primitive may become secondary to *selection*.

Using these interaction primitives, Samur introduces a taxonomy of haptic interaction (Samur, 2012), as shown in Fig. 1.7. This taxonomy is invaluable for designing new haptic systems by simplifying requirement derivation, identifying analogies for system components, and easing evaluation processes.

Other psychophysics-based approaches to haptic interaction include:

1. **Lederman and Klatzky**: They propose two classes of haptic operations: *Identification* (the "What-System") and *Localization* (the "Where-System") (Lederman & Klatzky, 2009).
2. **Hollins**: This approach categorizes primitives based on spatial and temporal resolution, alongside a class of "haptic" interactions that correspond to motor control primitives (Hollins, 2002).

Samur's taxonomy, as depicted in Fig. 1.7, offers a more straightforward application for developing task-specific haptic systems compared to the approaches of Lederman, Klatzky, and Hollins, making it the preferred choice for this text.

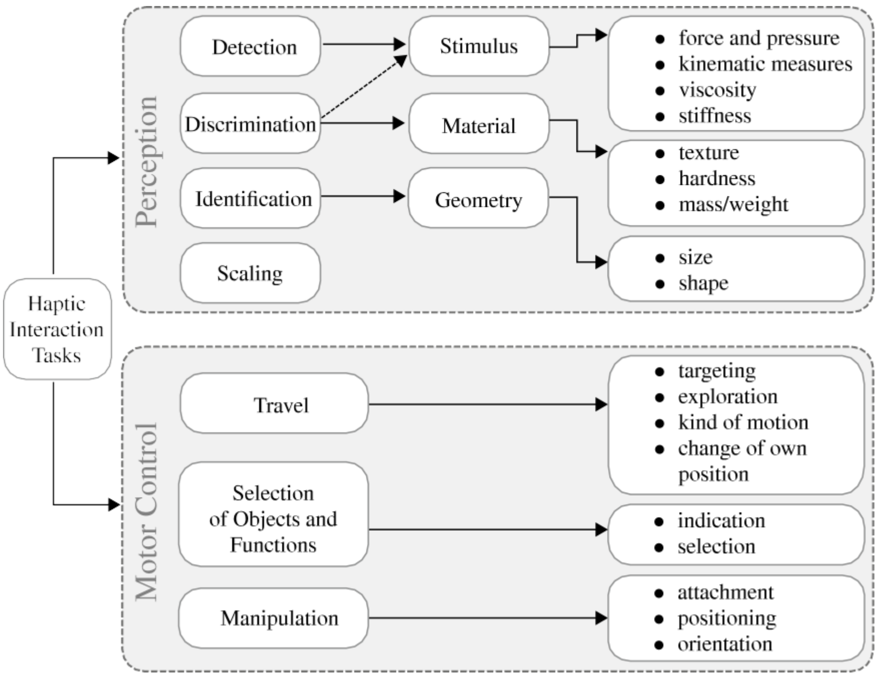


Figure 1.7 *Presents a taxonomy of haptic interaction, highlighting classifications derived from perception and motor control*

primitives. The figure is based on Gall et al. (2001) and Samur (2012).

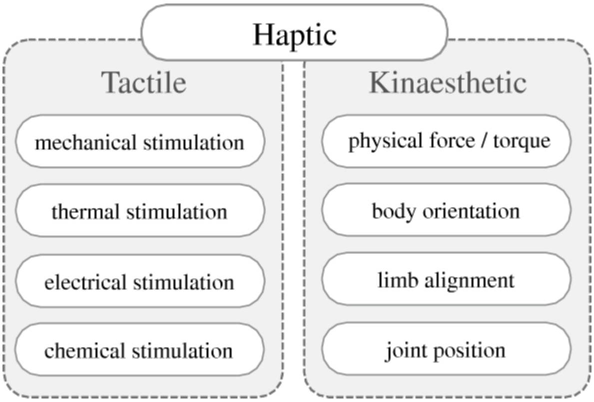


Figure 1.8 Illustrates the taxonomy of haptic perception, as defined in Kirkpatrick and Douglas (2002).

Taxonomy of Haptic Perception

One of the primary taxonomies in haptic literature that has not yet been discussed is the classification based on kinesthetic and tactile perception properties. This taxonomy is physiologically based, defining perception exclusively according to the location of sensory receptors. It is formally outlined in the ISO 9241-910 standard (Kirkpatrick & Douglas, 2002) and illustrated in Figure 1.8.

According to this definition, tactile perception involves all cutaneous receptors, which include not only mechanoreceptors but also receptors for temperature, chemicals (such as taste), and pain. While temperature and pain perception are significant, mechanical interaction is more practical for task-specific haptic systems due to its usability and general applicability. However, mechanical interaction also presents greater technical challenges due to the complexity and dynamics of mechanoreceptors. For this reason, the focus of this book is on mechanical perception and interaction.

For a detailed discussion of pain perception, readers are referred to specialized literature on the subject (Kruger et al., 1996). Although temperature perception and its potential applications are addressed in prior works (Darian-Smith & Johnson, 1977; Jones & Berris, 2002), thermal displays remain less critical than mechanical interaction in terms of information transfer and dynamics (Ino et al., 1993; Norrsell et al., 2001; Sato & Maeno, 2012).

Building on these insights, it becomes evident that mechanical stimuli play a central role in haptic system design due to their complexity and relevance to task-specific applications. Unlike thermal or pain-related feedback, mechanical interactions provide a richer and more versatile medium for conveying detailed information. This is particularly true in applications requiring precise feedback, such as surgical robotics, virtual reality environments, or advanced teleoperation systems. The intricate dynamics of mechanoreceptors and the associated challenges in replicating their response further underscore the need to prioritize mechanical stimuli in the development of haptic technologies.

With this focus on mechanical stimuli, kinesthetic and tactile perception can now be defined as follows:

Definition *kinaesthetic* kinaesthetic perception describes the perception of the operational state of the human locomotor system, particularly joint posi- tions, limb alignment, body orientation and muscle tension. For kinaesthetic perception, there are dedicated sensory receptors in muscles, tendons and joints as detailed in Sect. 2.1. Regarding the taxonomy of haptic interactions, kinaes- thetic sensing is primarily involved the motion control primitives, since signals from kinaesthetic receptors are needed in the biological control loop for

Definition *tactile* Tactile perception describes the perception based on sen- sory receptors located in the human skin. Compared to kinaesthetic receptors, they exhibit much larger dynamics and are primarily involved in the perception primitives of haptic interaction.

Historically, the terms *tactile* and *kinesthetic* have been strictly defined based on the location and function of sensory receptors. Recently, however, their usage has broadened. Although the root of *kinesthesia* relates to movement, the term *kinesthetic* is now also applied to describe static conditions (Loomis & Lederman, 1986). Additionally, *kinesthetic* is sometimes used to refer specifically to limb-related perception, while *proprioception* describes the perception of the whole body (Clark & Horch, 1986). However, this distinction is not emphasized further in this book due to its limited technical significance. Similarly, *tactile* is often used generically to describe sensors or actuators with spatial resolution, even when not addressing tactile perception as traditionally defined.

The definition of *tactile* and *kinesthetic* has also evolved to include their dynamic properties, extending their relevance to haptic interactions. Based on Shimoga's work, kinesthetic dynamics align with the motion capabilities of the locomotor system (Shimoga, 1993). Tactile dynamics are practically limited to frequencies between 1–2 kHz, as higher frequencies, while perceivable (Gault, 1927; Verrillo & Gescheider, 1992), do not significantly contribute to perception (Brooks, 1990). As discussed in Section 2.4.3, this limitation is necessary for the design of electromechanical components in haptic systems (Fig. 1.9).

Extending these dynamics, Daniel and McAree (1998) proposed a bidirectional, asymmetric model with a low-frequency (<30 Hz) channel for energy exchange and a high-frequency channel for information exchange. This mapping of dynamic properties is critical because users, above the active movement dynamics of the locomotor system, can be treated as passive mechanical systems (Hogan, 1989). Together, these concepts form the widely accepted model for partitioning haptic interaction into low-frequency kinesthetic interaction and high-frequency tactile perception.

The classification of haptic interaction and perception, along with their dynamic models, has established a comprehensive framework for understanding and designing haptic systems. These taxonomies highlight the intricate relationships between tactile and kinesthetic modalities, offering a structured approach to capturing the complexity of human sensory experiences. By defining interaction dynamics and perceptual boundaries, this framework provides a foundation for creating systems that closely mimic or enhance human haptic capabilities, ensuring their functionality aligns with intended use cases.

The introduction of standardized descriptions for haptic interaction simplifies the design process by offering a common language for researchers and engineers. This standardization not only facilitates collaboration but also enhances the clarity of system functionalities, ensuring that task-specific designs meet the precise requirements of their intended applications. The ability to clearly communicate design intentions and performance criteria is particularly crucial in fields such as robotics, virtual reality, and medical simulation, where haptic feedback is a key component of the user experience.

Dynamic models of haptic perception further refine this framework by integrating the physical limitations and sensory thresholds of human interaction. These models guide the design of haptic systems by delineating the roles of low-frequency and high-frequency feedback in different applications. By focusing on these aspects, designers can develop systems that optimize user interaction, enhance sensory feedback, and improve overall efficiency. Such models ensure that haptic systems are both technically robust and user-centric, paving the way for advancements in various domains requiring precise and immersive interaction capabilities.

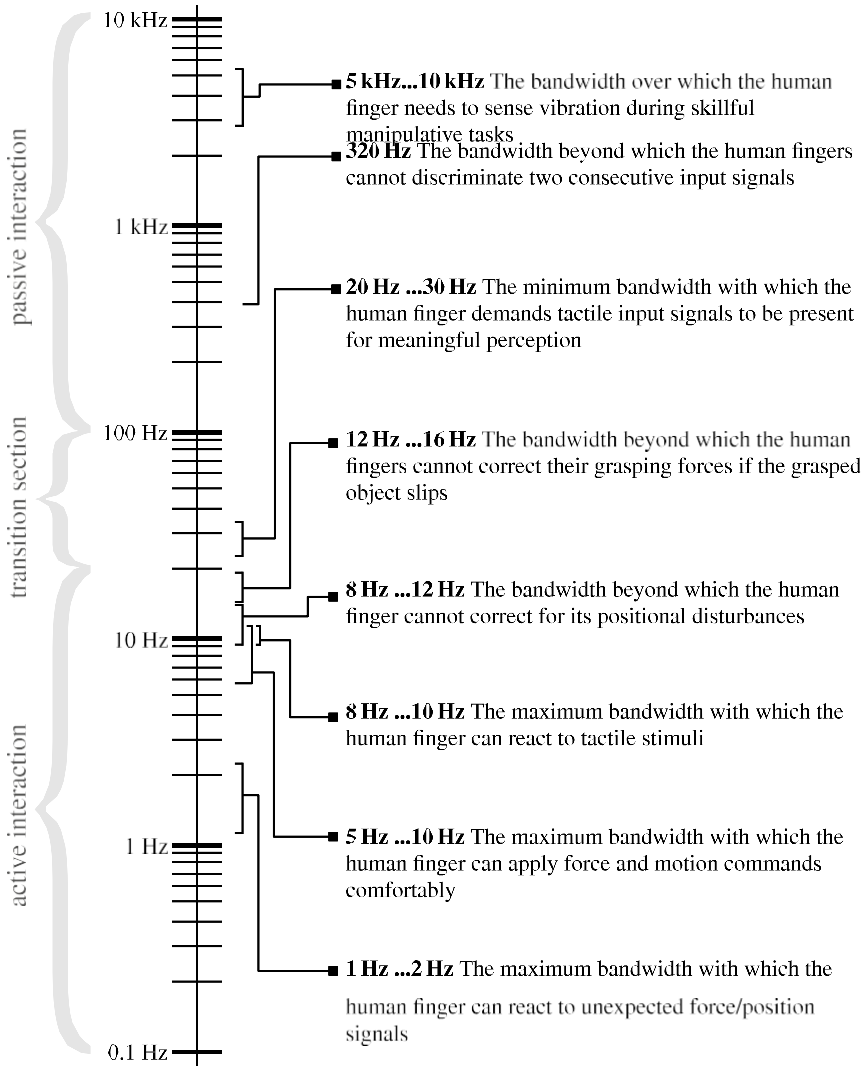


***Figure 1.9** illustrates kinesthetic and tactile haptic interaction, based on data from Gault (1927), Brooks (1990), and Shimoga (1993).*

Application Areas of Haptic Systems

Haptic systems find applications in diverse fields, each leveraging the unique capabilities of tactile and kinaesthetic feedback to enhance user interaction. These systems offer significant advantages, such as improving task efficiency, enabling realistic simulations, and providing intuitive interfaces. However, they also present technical challenges, including the complexity of designing responsive and reliable feedback mechanisms. Exploring the benefits and addressing the hurdles associated with haptic systems contributes to the development of robust solutions tailored to specific user needs and scenarios.

Telepresence, Teleaction and Assistive Systems

Have you ever thought about touching a lion in a zoo's cage? With a telepresence and teleaction (TPTA) system, such interactions are possible without any personal risk. TPTA systems allow mechanical interaction with remote environments by converting mechanical signals into other domains, such as electrical, making transmission feasible. These systems typically feature multimodal capabilities, such as a unidirectional visual channel displaying the environment to the operator.

Applications of TPTA systems include underwater assembly, where visual cues are limited due to dispersed particles (Dennerlein et al., 1997), scaled micro- and nano-positioning (Estevez et al., 2010; Vander Poorten et al., 2012), and surgical applications (Westebring-van der Putten et al., 2008; Wagner et al., 2002). In surgery, they enable innovative techniques like palpation in minimally invasive procedures, improving task efficiency, reducing errors, and enhancing patient safety (Nitsch & Färber, 2012; Wagner et al., 2002). The increasing bandwidth of networked systems, especially with 5G, is driving new possibilities (Antonakoglou et al., 2018). However, effective input devices remain critical for system usability (Young & Peschel, 2020).

Most TPTA systems are research-focused. For example, Quanser offers a haptic interface integrated with a robotic manipulator, enabling versatile bilateral teleoperation scenarios. Systems like neuroArm facilitate neurological interventions requiring high precision and real-time MRI integration (Sutherland et al., 2013).

Developing TPTA systems is technically challenging due to unknown environmental properties affecting system stability, the high accuracy needed for sensors and actuators, and the complexity of transmitting data over long distances with potential packet loss and latency issues.

A specific subtype of TPTA systems, comanipulators, is primarily used in medical applications (Westebring-van der Putten et al., 2008). These systems not only enable mechanical interaction but also provide additional manipulation and feedback. Examples include INKOMAN and HapCath, developed at the Institute for Electromechanical Design.

Figure 1.10 illustrates versatile teleoperation using Quanser's HD2 haptic interface, which provides 7 degrees of freedom (DoF) of haptic feedback, paired with the Denso Open Architecture robot offering 6 DoF. Image courtesy of Quanser, Markham, Ontario, Canada, used with permission.

The HapCath system, which introduces haptic feedback for cardiovascular interventions, is detailed in Section 14.2. Figure 1.11 illustrates the INKOMAN instrument, developed through the SOMIT-FUSION research project funded by the German Ministry of Education and Research. This instrument extends a laparoscopic device with a parallel kinematic structure (Röse, 2011) to provide additional degrees of freedom (DOF) on a universal tool platform (Schlaak et al., 2008). This enhancement enables minimally invasive procedures in previously inaccessible regions of the liver. By incorporating a multi-component force sensor (Rausch, 2012), the platform allows the user to perceive interaction forces between the instrument and the liver (Kassner, 2013), facilitating techniques like palpation to detect vessels or cancerous tissues. With its laparoscopic form, the instrument also permits additional interaction forces to be applied by the surgeon, classifying it as a comanipulation system.

TPTA systems remain a primary focus of research due to their niche markets with high potential. Medical applications, however, are an exception, where indirectly coupled instruments enhance safety and efficiency, avoiding collisions between tools and reducing contact or grip forces (Nitsch & Färber, 2012; Wagner et al., 2002). Additionally, automated tasks like knot tying can be performed faster and more reliably (Bauernschmitt et al., 2005).

The distinction between haptic TPTA systems and robotic medical systems can be subtle, as many robotic functions do not rely on haptic feedback. This explains the prevalence of robotic systems in research and industry, exemplified by the Da Vinci system from Intuitive Surgical Operations Inc. (Pott et al., 2005). Designed for urological and gynecological procedures, Da Vinci features a handling console with a 3D view of the surgical site and a variety of cable-driven instruments operated by the surgeon (Guthart & Salisbury, 2000).

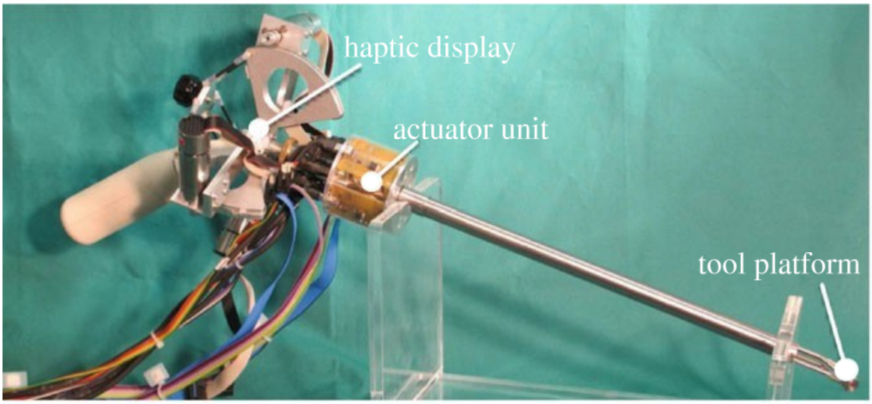


Figure 1.11 *illustrates INKOMAN, an intracorporal manipulator designed for minimally invasive abdominal interventions, offering enhanced flexibility. The handheld instrument includes a haptic display based on a delta kinematic structure. The tool platform's parallel kinematic structure is powered by ultrasonic traveling wave motors. Adapted from Kassner (2013).*

Figure 1.12 depicts the Da Vinci SP surgical system designed for single-port access. Image courtesy of Intuitive Surgical Operations, Inc., Qc 2022, used with permission.

Holland Haptics developed a consumer product called Frebble, designed to simulate the sensation of holding someone's hand over the internet. This device was not only an intriguing hardware concept but also served as a low-cost teleoperation tool.

Significant advancements have also been made in magnetic resonance imaging (MRI) studies on hand neural control. However, the harsh MRI environment presents challenges for devices capable of delivering diverse stimuli. One study introduced an fMRI-compatible haptic interface designed to investigate the neural mechanisms of precision grasp control. The system, positioned at the scanner bore, employs a shielded electromagnetic actuation system located at the scanner bed's end and utilizes a high-stiffness cable. Performance testing demonstrated renderable forces up to 94 N, structural stiffness of 3.3 N/mm, and a position control bandwidth of at least 19 Hz.

This system actuates two degrees of freedom (DOF) per finger using closed-loop cable transmissions supported by aluminum profiles with redirection modules. The cables pass through mechanisms for length and tension adjustment and are guided by low-friction pulleys with polymer/glass ball bearings. These components are mounted on an aluminum bar rigidly attached to the scanner bedside. To prevent slippage, cables are fixed to the capstan. Due to transmission friction and potential cable wear, the design allows for quick cable replacement during fMRI studies to maintain seamless interaction with operators.

Virtual Environments

A significant application area for haptic systems is their use in virtual environments. Given the breadth of this field, specific areas where interactions with simulated scenarios are extensively applied are outlined below:

Medical Training Numerous haptic systems are designed for medical training, enabling realistic practice without risking harm to actual patients (Coles et al., 2010). These systems often combine haptic feedback with visual and acoustic inputs to create an immersive simulation of medical procedures. Examples include simulators for diagnosing joint lesions (Bajka et al., 2009) and training systems for endoscopic, laparoscopic, and intravascular interventions (Samur, 2012). Figure 1.13 provides an example of a surgical simulator. Studies show that surgeons trained on simulators demonstrate improved task performance (Ahlberg et al., 2007; Bajka et al., 2009). Additionally, simulators can be introduced early in training, offering patient safety and greater accessibility.

Industrial Design In industrial design, haptic systems simulate assembly operations and enable subjective evaluation of prototypes. Although fewer applications exist compared to medical training, this field drives technological advancement. For instance, unique requirements have led to the development of admittance systems and form displays, such as the Haptic Strip. This device features a bendable, twistable surface with six degrees of freedom (6 DoF), allowing designers to evaluate large-scale forms without creating physical prototypes (Bordegoni et al., 2010). Figure 1.14 illustrates the Haptic Strip's functionality.

Multimodal Information Displays Since the haptic sense is naturally suited to analyzing objects and environments, it is also applied in fields requiring intuitive access to complex information. Haptic systems are used in biology and chemistry to visualize large datasets (Burdea & Coiffet, 2003) and to support molecular synthesis (Brown et al., 2006). For these applications, the human ability to recognize patterns in visual representations facilitates coarse positioning, while haptic feedback assists with precise micro-positioning by simulating intermolecular forces. A recent example of a multimodal information display was introduced by Microsoft Research (Sinclair et al., 2013)

Figure 1.13 Depicts the LAP Mentor III laparoscopic simulator, designed to replicate abdominal interventions. Image courtesy of Simbionix USA,Cleveland, OH, USA, used with permission.

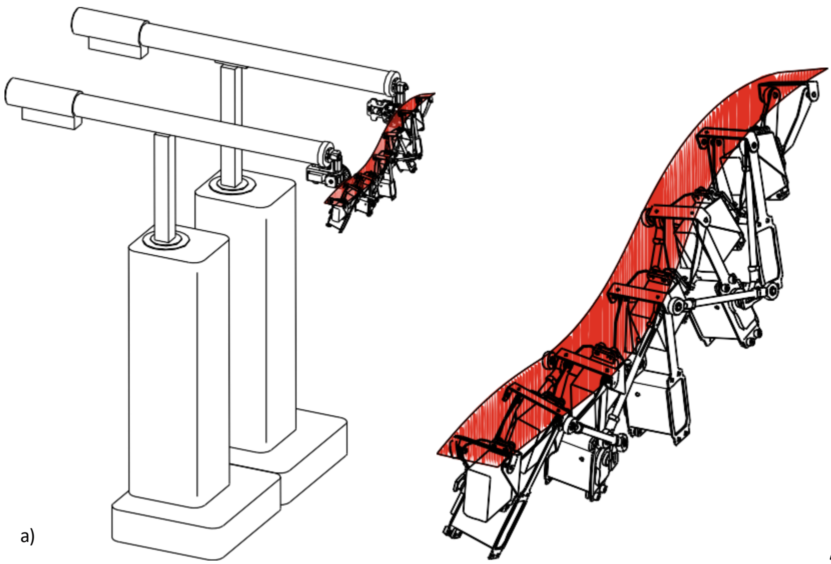
The TouchMover is an actuated screen featuring haptic feedback, enabling users to interact with object and material properties or intuitively explore volumetric data, such as MRI scans. Figure 1.15 demonstrates this application, where annotations are highlighted both visually and haptically with detents, promoting intuitive access and collaboration.

Consumer Electronics In the realm of computer games, haptic feedback integration has seen varied success. Novint Technologies introduced the Falcon haptic interface in 2006, which features a delta parallel kinematic structure and a competitive price of approximately $500. Due to its affordability and compatibility with various application programming interfaces (APIs), the Falcon has also been adopted in several research projects (Shah et al., 2010).

From a 202x perspective, complex haptic devices have struggled to establish themselves in consumer electronics. Currently, their primary application is limited to gamepads and controllers that offer vibrotactile feedback. However, Sony’s DualSense

Technology, which combines vibration actuators with motorized adaptive triggers, has increased the complexity of haptic-enabled input devices. Whether this signifies a resurgence of kinesthetic feedback in consumer electronics remains to be seen.

Other applications of haptic technology exist. For example, systems like Butt Kicker by The Guitammer Company enhance gaming immersion by conveying low-frequency acoustic signals (Figure 1.16). Additionally, the Haptex project has developed rendering algorithms and hardware interfaces to facilitate the tactile experience of fabrics over the internet.



a) b)

Figure 1.14 *illustrates the Haptic Strip system, mounted on two HapticMaster admittance-type interfaces. Capacitive sensors embedded in the strip surface detect user touch. Adapted from Bordegoni et al. (2010), Qc Springer Nature, all rights reserved.*

Figure 1.15 *Shows the TouchMover system in use, with a user exploring MRI data. Image courtesy of Microsoft Research, Redmond, WA, USA, used with permission.*

Compared to the design of TPTA systems, developing haptic interfaces for interactions with virtual environments is somewhat less complex, as the interaction environment is better understood during the design process. However, challenges such as deriving and allocating environmental data emerge in these applications. Additionally, due to the broader adoption of such systems, cost efficiency becomes a critical factor.

Figure 1.16 *Shows the ButtKicker electrodynamic actuator, designed to generate low-frequency oscillations on a gaming seat. Image courtesy of The Guitammer Company, Qc 2022, used with permission.*

Non-invasive Medical Applications

Haptic perception parameters can be instrumental in diagnosing certain illnesses and dysfunctions. For instance, diminished haptic perception is associated with conditions like eating disorders (Grunwald et al., 2001; Grunwald, 2008) and diabetic neuropathy (Norrsell et al., 2001). Diagnosis is achieved by measuring perception or motor exertion parameters and comparing them to population norms. Beyond diagnosis, these parameters also serve as progress indicators in rehabilitation for stroke (Allin et al., 2002) and limb recovery (Yang et al., 2007).

To support such applications, cost-effective systems with robust and efficient measurement protocols are essential. Unlike TPTA or VR systems, user feedback can be collected with simpler methods, making development comparatively less complex. Although several research groups are actively working in this field, no comprehensive solution has yet reached the market.

Communication

The most widely used haptic systems are in basic communication applications, such as the vibration function on mobile phones. Unlike visual or auditory signals, haptic communication enables discrete information transfer with spatial resolution, making it intuitive—feedback occurs directly at the point of interaction. A common example is a switch that provides haptic feedback when pressed.

Haptic communication is particularly beneficial in demanding environments, such as driving. Studies show it distracts users less than visual or auditory channels (Ryu et al., 2010; Spence & Ho, 2008). Applications include military navigation systems

(Gilson et al., 2007) and adaptive automotive interfaces for raising driver awareness in potentially hazardous situations (Liedecke et al., 2014). The shift towards steer-by-wire systems and autonomous vehicles underscores the importance of the haptic channel in these applications.

The growing prevalence of touchscreens in consumer electronics has driven demand for haptic feedback technologies to simplify usage without requiring visual inspection. These solutions employ various actuation principles, which will be detailed in Chapter 9.

Haptic systems also benefit visually impaired users by enabling tactile interfaces for Braille characters, navigation aids, and interaction with graphic interfaces. Projects like HaptiMap provide toolkits for standard mobile devices (HaptiMap, 2007), while studies show vibrotactile feedback can support stroke rehabilitation (Seim et al., 2021). **Figure 1.17** illustrates examples of haptic systems in communication.

Shape-changing interfaces are another innovative application. By altering their form, these devices provide navigation assistance, particularly for visually impaired users, by guiding them through tactile sensations. The system utilizes a bi-directional expanding mechanism driven by a motor with a rack-and-pinion system for translational movement. The ergonomic design allows the device to rest comfortably in the palm.

Scientific research in this area also explores haptic icons, or "tactons," for information transfer. Variables such as rhythm, signal form, frequency, and localization are studied (Brown et al., 2006; Enriquez et al., 2006). Current transfer rates range from 2 to 12 bits per second, with specialized devices like the Tactuator enabling higher rates (Cholewiak et al., 2008). However, the exact bandwidth for practical use remains unclear. A study by Seo and Choi (2015) reported a transfer rate of 3.7 bits per second.

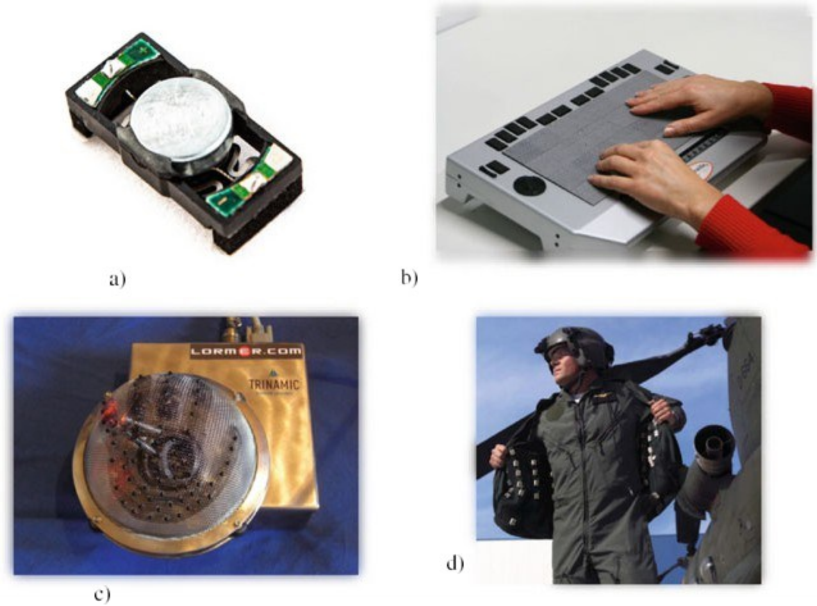


Figure 1.17 *Showcases various components and systems for haptic communication:*
a) Grewus Exciter EXR4403L-01A for touchpads and mobile devices.
b) Hyperbraille system for displaying graphic information for visually impaired users (courtesy of metec AG, Stuttgart, Germany). c) Lormer system, a machine-human interface using the Lorm alphabet on the palm and hand to convey text (courtesy of Thomas Rupp). d) Tactile Torso Display, a vest designed to display flight information on a pilot's torso (courtesy of TNO, Soesterberg, The Netherlands).
All images used with permission.

Completing the Picture

Passive systems like computer keyboards, trackballs, and mice also fall within the scope of haptic communication, as they convey motion control operations to computer systems. Although they provide some degree of haptic feedback, this feedback is independent of user interaction and instead relies on the system's physical properties, such as inertia, damping, or friction.

Another exciting area influenced by haptic research is robotic hands and limbs equipped with sensors inspired by human perception. Tactile sensors, a core aspect of haptic research play a pivotal role in robotics particularly in bionics-inspired systems (Bianchi & Moscatelli, 2016). A milestone in this field is the micromechanical design of a fully dexterous robotic hand integrated with advanced capacitive pressure sensors (Figure 1.19). The potential of these systems extends far beyond humanoid applications, offering promising innovations across diverse domains.

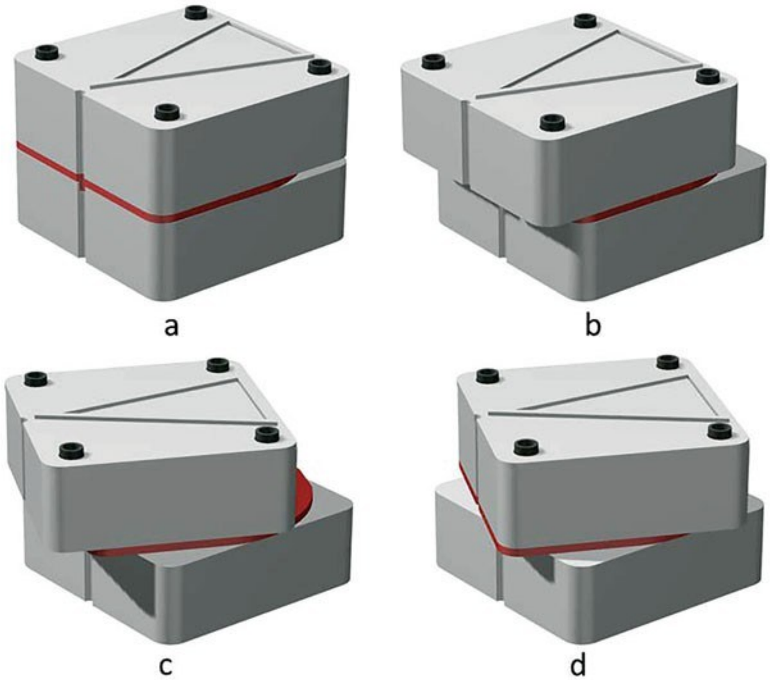


***Figure 1.18** Illustrates various shapes of a haptic interface designed to send different commands. Images courtesy of Ad Spiers, used with permission, based on Spiers and Dollar (2016).*

Figure 1.19 showcases the fully actuated Shadow Dexterous Hand by Shadow Robot Company, integrated with BioTacs sensors from SynTouch. This system enables precise manipulation with direct contact force and direction measurement for each fingertip. Image courtesy of Shadow Robot Company, Qc 2022, used with permission.

Why Use a Haptic System?

There are numerous reasons to utilize a haptic system: improving task performance, reducing error rates in manipulation scenarios, utilizing an untapped sensory channel to convey additional information, or gaining a competitive edge in innovation-driven markets. While this book does not aim to evaluate whether haptics can meet these expectations, it focuses on the design of specific haptic systems tailored for intended applications.

Although guidelines exist for implementing haptic and multimodal feedback to optimize task performance there are limited resources to determine the usability of haptic feedback for a given application. Acker (2011) outlines criteria for telepresence technologies in industrial settings, while Jones and Sarter (2008) provide guidelines for tactile system usage.

Conclusions

Haptic systems play a pivotal role across a wide range of applications, enabling intuitive and efficient user interactions. This book lays the groundwork for understanding and designing task-specific haptic interfaces by exploring the fundamental principles of haptic interaction and the user's role within these systems. Building on this foundation, it provides detailed insights into the development and structural considerations essential for creating effective haptic systems. With these concepts established, the subsequent chapters transition into the practical aspects of designing tailored haptic solutions, ensuring a comprehensive understanding of both theoretical and applied perspectives.

References

Acker, A. (2011). *Anwendungspotential von Telepräsenz- und Teleaktionssystemen für die Präzisionsmontage* (Doktora tezi). Technische Universität München. http://mediatum.ub.tum.de/doc/1007163/1007163.pdf

Adams, R. J. (1999). Stable haptic interaction with virtual environments (PhD thesis). University of Washington, Washington.

Anuguelov, N. (2009). Haptische und akustische Kenngrößen zur Objektivierung und Optimierung der Wertanmutung von Schaltern und Bedienfeldern für den KFZ-Innenraum (Doktora tezi). Technische Universität Dresden. http://d-nb.info/1007551739/

Reisinger, J. (2009). Parametrisierung der Haptik von handbetätigten Stellteilen (Doktora tezi). Technische Universität Münhttps://mediatum.ub.tum.de/doc/654165/654165.pdf

Ahlberg, G., et al. (2007). Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. American Journal of Surgery, 193(6), 797–804. https://doi.org/10.1016/j.amjsurg.2006.06.050

Ahlberg, G., et al. (2007). Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. *American Journal of Surgery,* *193*(6), 797–804. https://doi.org/10.1016/j.amjsurg.2006.06.050

Antonakoglou, K., et al. (2018). Toward haptic communications over the 5G tactile internet. IEEE Communications

Surveys & Tutorials, 20(4), 3034–3039. https://doi.org/10.1109/COMST.2018.2851452

Bajka, M., et al. (2009). Evaluation of a new virtual-reality training simulator for hysteroscopy. Surgical Endoscopy, 23(9), 2026–2033. https://doi.org/10.1007/s00464-008-9927-7

Bajka, M., et al. (2009). Evaluation of a new virtual-reality training simulator for hysteroscopy. *Surgical Endoscopy, 23*(9), 2026–2033. https://doi.org/10.1007/s00464-008-9927-7

Bauernschmitt, R., et al. (2005). Towards robotic heart surgery: Introduction of autonomous procedures into an experimental surgical telemanipulator system. International Journal of Medical Robotics and Computer Assisted Surgery, 1(3), 74–79. https://doi.org/10.1002/rcs.30

Bianchi, M., & Moscatelli, A. (2016). Human and robot hands. Springer. https://doi.org/10.1007/978-3-319-26706-7

Bordegoni, M., et al. (2010). A force and touch sensitive self-deformable haptic strip for exploration and deformation of digital surfaces. In Kappers, A. M. L., Bergmann Tiest, W. M., & van der Helm, F. C. T. (Eds.), Haptics: Generating and Perceiving Tangible Sensations (pp. 65–72). Heidelberg: Springer. https://doi.org/10.1007/978-3-642-14075-4_10

Bordegoni, M., et al. (2010). A force and touch sensitive self-deformable haptic strip for exploration and deformation of digital surfaces. In Kappers, A. M. L., Bergmann Tiest, W. M., & van der Helm, F. C. T. (Eds.), *Haptics: Generating and Perceiving Tangible Sensations* (pp. 65–72). Springer. https://doi.org/10.1007/978-3-642-14075-4_10

Bowman, D., & Hodges, L. (1999). Formalizing the design, evaluation, and application of interaction techniques for immersive virtual environments. Journal of Visual Languages and Computing, 10(1), 37–53. https://doi.org/10.1006/jvlc.1998.0111

Breitschaft, S. J., & Carbon, C. C. (2021). Function follows form: Using the aesthetic association principle to enhance haptic interface design. *Frontiers in Psychology, 12*, 2619. https://doi.org/10.3389/fpsyg.2021.646986

Brown, L., Brewster, S., & Purchase, H. (2006). Multidimensional tactons for non-visual information presentation in mobile devices. In Proceedings of the Conference on Human-Computer Interaction with Mobile Devices and Services (pp. 231–238). https://doi.org/10.1145/1152215.1152265

Burdea, G. C., & Coiffet, P. (2003). Virtual Reality Technology (2nd ed.). Wiley-Interscience. ISBN: 978-0471360896

Burdea, G. C., & Coiffet, P. (2003). Virtual Reality Technology. Wiley-Interscience. ISBN: 978-0471360896

Carbon, C.-C., & Jakesch, M. (2013). A model for haptic aesthetic processing and its implications for design. Proceedings of the IEEE, 101(9), 2123–2133. https://doi.org/10.1109/JPROC.2012.2219831

Cholewiak, S. A., Tan, H. Z., & Ebert, D. S. (2008). Haptic identification of stiffness and force magnitude. In Symposium on Haptic Interfaces for Virtual Environments and Teleoperator Systems (pp. 129–135). Reno. https://doi.org/10.1109/HAPTICS.2008.4479918

Clark, F., & Horch, K. (1986). Kinesthesia. In Boff, K. R., Kaufman, L., & Thomas, J. P. (Eds.), Handbook of Perception and Human Performance (pp. 13.1–13.61). New York: Wiley-Interscience. ISBN: 978-0471829577

Lederman, S. J., & Klatzky, R. L. (2009). Haptic perception: A tutorial. In Attention, Perception, & Psychophysics, 71(7), 1439–1449. https://doi.org/10.3758/APP.71.7.1439

Coles, T. R., Meglan, D., & John, N. W. (2010). The role of haptics in medical training simulators: A survey of the state-of-the-art. IEEE Transactions on Haptics, 4(1), 51–66. https://doi.org/10.1109/TOH.2010.19

Coles, T. R., Meglan, D., & John, N. W. (2010). The role of haptics in medical training simulators: A survey of the state-of-the-art. *IEEE Transactions on Haptics, 4*(1), 51–66. https://doi.org/10.1109/TOH.2010.19

Crusco, A. H., & Wetzel, C. G. (1984). The Midas touch: The effects of interpersonal touch on restaurant tipping. Personality and Social Psychology Bulletin, 10(4), 512–517. https://doi.org/10.1177/0146167284104003

Culbertson, H., & Kuchenbecker, K. J. (2017). Importance of matching physical friction, hardness, and texture in creating realistic haptic virtual surfaces. IEEE Transactions on Haptics, 10(1), 63–74. https://doi.org/10.1109/TOH.2016.2598751

Dennerlein, J., Millman, P., & Howe, R. (1997). Vibrotactile feedback for industrial telemanipulators. In Sixth Annual Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. http://biorobotics.harvard.edu/pubs/1997/vibrotactile.pdf

Enriquez, M., MacLean, K., & Chita, C. (2006). Haptic phonemes: Basic building blocks of haptic communication. In Proceedings of the 8th International Conference on Multimodal Interfaces (ICMI) (pp. 302–309). ACM. https://doi.org/10.1145/1180995.1181053

Estevez, P., et al. (2010). A haptic tele-operated system for microassembly. In Ratchev, S. (Ed.), Precision Assembly Technologies and Systems (pp. 13–20). Heidelberg: Springer. https://doi.org/10.1007/978-3-642-11598-1_2

Fiene, J., Kuchenbecker, K. J., & Niemeyer, G. (2006). Event-based haptic tapping with grip force compensation. In IEEE Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. https://doi.org/10.1109/HAPTIC.2006.1627063

Gaspar, J., et al. (2019). Push button design requirements and relations to button architecture elements. International Journal of Industrial Ergonomics, 70, 92–106. https://doi.org/10.1016/j.ergon.2019.01.001

Gerling, G., & Thomas, G. (2005). The effect of fingertip microstructures on tactile edge perception. In First Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (pp. 63–72). https://doi.org/10.1109/WHC.2005.129

Grunwald, M., et al. (2001). Haptic perception in anorexia nervosa before and after weight gain. *Journal of Clinical and Experimental Neuropsychology, 23*(4), 520–529. https://doi.org/10.1076/jcen.23.4.520.1229

Guthart, G., & Salisbury, J. Jr. (2000). The intuitive telesurgery system: Overview and application. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (Vol. 1, pp. 618–621). San Francisco. https://doi.org/10.1109/ROBOT.2000.844121

Haptex. (2007). Haptex project: Haptic rendering algorithms for the sense of touch. European Union IST-6549. http://haptex.miralab.unige.ch/

HaptiMap. (2007). *Haptic, audio and visual interfaces for maps and location-based services*. European Union FP7-ICT-224675. http://www.haptimap.org

HaptiMap. (2007). Haptic, audio, and visual interfaces for maps and location-based services. European Union FP7-ICT-224675. http://www.haptimap.org

Hollins, M. (2002). Touch and haptics. In Pashler, H. (Ed.), Steven's Handbook of Experimental Psychology (pp. 585–618). New York: Wiley. ISBN: 978-0471377771. https://doi.org/10.1002/0471214426.pas0114

25. Kruger, L., Friedman, M. P., & Carterette, E. C. (Eds.). (1996). Pain and touch. Maryland Heights, MO: Academic Press. ISBN: 978-0123992390

Hollins, M. (2002). Touch and haptics. In Pashler, H. (Ed.), Steven's Handbook of Experimental Psychology (2nd ed., pp. 585–618). New York: Wiley. ISBN: 978-0471377771

Israr, A., et al. (2019). Stereohaptics toolkit for dynamic tactile experiences. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11786, 217–232. https://doi.org/10.1007/978-3-030-30033-3_17

Jagodzinski, R., & Wintergerst, G. (2009). A toolkit for developing haptic interfaces - control dial based on electromagnetic brake. In Proceedings of UIST (pp. 184). Victoria, Canada. http://www.acm.org/uist/archive/adjunct/2009/pdf/demos/paper184.pdf

Johansson, G. (2008). Haptics as a substitute for vision. In Hersh, M. A., & Johnson, M. A. (Eds.), *Assistive Technology for Visually Impaired and Blind People* (pp. 135–166). Springer. ISBN: 978-1846288661

Johansson, G. (2008). Haptics as a substitute for vision. In Hersh, M. A., & Johnson, M. A. (Eds.), Assistive Technology for Visually Impaired and Blind People (pp. 135–166). Springer. https://doi.org/10.1007/978-1-84628-866-1

Jones, L. A., & Sarter, N. B. (2008). Tactile displays: Guidance for their design and application. Human Factors, 50(1), 90–111. https://doi.org/10.1518/001872008X250638

Kassner, S. (2013). Haptische Mensch-Maschine-Schnittstelle für ein laparoskopisches Chirurgie-System (Doktora tezi). Technische Universität Darmstadt. http://tubiblio.ulb.tu-darmstadt.de/63334/

Kassner, S. (2013). Haptische Mensch-Maschine-Schnittstelle für ein laparoskopisches Chirurgie-System (Doktora tezi). Technische Universität Darmstadt. http://tubiblio.ulb.tu-darmstadt.de/63334/

King, C.-H., et al. (2009). Tactile feedback induces reduced grasping force in robot-assisted surgery. IEEE Transactions on Haptics, 2(2), 103–110. https://doi.org/10.1109/TOH.2009.4

Kruger, L., Friedman, M. P., & Carterette, E. C. (Eds.). (1996). Pain and Touch. Academic Press. ISBN: 978-0123992390

Lederman, S. J., & Klatzky, R. L. (2009). Haptic perception: A tutorial. Attention, Perception, & Psychophysics, 71(7), 1439–1459. https://doi.org/10.3758/APP.71.7.1439

Nitsch, V., & Färber, B. (2012). A meta-analysis of the effects of haptic interfaces on task performance with teleoperation systems. IEEE Transactions on Haptics, 6(4), 387–398. https://doi.org/10.1109/TOH.2012.62

Okamoto, S., Nagano, H., & Yamada, Y. (2013). Psychophysical dimensions of tactile perception of textures. IEEE Transactions on Haptics, 6(1), 81–93. https://doi.org/10.1109/TOH.2012.32

Papetti, S., & Saitis, C. (2018). Musical haptics. Springer. https://doi.org/10.1007/978-3-319-58316-7

Papetti, S., & Saitis, C. (Eds.). (2018). Musical Haptics. Springer Series on Touch and Haptic Systems. https://doi.org/10.1007/978-3-319-58316-7

Riener, R., et al. (2002). Orthopädischer Trainingssimulator mit haptischem Feedback. Automatisierungstechnik, 50(6), 296–301. https://doi.org/10.1524/auto.2002.50.6.296

Riener, R., et al. (2002). Orthopädischer Trainingssimulator mit haptischem Feedback. *Automatisierungstechnik, 50*(6), 296–301. https://doi.org/10.1524/auto.2002.50.6.296

Röse, A. (2011). Parallelkinematische Mechanismen zum intrakorporalen Einsatz in der laparoskopischen Chirurgie (Doktora tezi). Technische Universität Darmstadt. http://tuprints.ulb.tu-darmstadt.de/2493/

Ryu, J., et al. (2010). Vibrotactile feedback for information delivery in the vehicle. IEEE Transactions on Haptics, 3(2), 138–149. https://doi.org/10.1109/TOH.2010.1

Samur, E. (2012). Performance metrics for haptic interfaces. Heidelberg: Springer. ISBN: 978-1447142249. https://doi.org/10.1007/978-1-4471-4225-6

Sato, K., & Maeno, T. (2012). Presentation of sudden temperature change using spatially divided warm and cool stimuli. In Kappers, A. M. L., Bergmann Tiest, W. M., & van der Helm, F. C. T. (Eds.), Haptics: Perception, Devices, Mobility, and Communication (pp. 457–468). Heidelberg: Springer. https://doi.org/10.1007/978-3-642-31401-8_41

Schlaak, H. F., et al. (2008). A novel laparoscopic instrument with multiple degrees of freedom and intuitive control. In Antwerpen, B. (Ed.), 4th European Conference of the International Federation for Medical and Biological Engineering (pp. 399–402). https://doi.org/10.1007/978-3-540-89208-3_394

Seim, C. E., Wolf, S. L., & Starner, T. E. (2021). Wearable vibrotactile stimulation for upper extremity rehabilitation in chronic stroke: Clinical feasibility trial using the VTS Glove. Journal of NeuroEngineering and Rehabilitation, 18(1), 1–11. https://doi.org/10.1186/s12984-021-00813-7

Seo, J., & Choi, S. (2015). Edge flows: Improving information transmission in mobile devices using two-dimensional vibrotactile flows. In IEEE World Haptics Conference (WHC) (pp. 25–30). https://doi.org/10.1109/WHC.2015.7177686

Sinclair, M., Pahud, M., & Benko, H. (2013). Actuated 3D touchscreen with haptic feedback. Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces (pp. 287–296). https://doi.org/10.1145/2512349.2512805

Sinclair, M., Pahud, M., & Benko, H. (2013). TouchMover: Actuated 3D touchscreen with haptic feedback. In Proceedings of the 2013 ACM International Conference on Interactive Tabletops and Surfaces (pp. 287–296). ACM. https://doi.org/10.1145/2512349.2512805

Sinclair, M., Pahud, M., & Benko, H. (2013). TouchMover: Actuated 3D touchscreen with haptic feedback. In *Proceedings of the 2013 ACM International Conference on Interactive Tabletops and Surfaces* (pp. 287–296). ACM. https://doi.org/10.1145/2512349.2512805

Spiers, A. J., & Dollar, A. M. (2016). Design and evaluation of shape-changing haptic interfaces for pedestrian navigation assistance. IEEE Transactions on Haptics, 10(1), 17–28. https://doi.org/10.1109/TOH.2015.2480680

Stevens, S. S. (1975). Psychophysics. Piscataway, NJ: Transaction Books. ISBN: 978-088738643522. Bowman, D., & Hodges, L. (1999). Formalizing the design, evaluation, and application of interaction techniques for immersive virtual environments. Journal of Visual Languages and Computing, 10(1), 37–53. https://doi.org/10.1006/jvlc.1998.0111

Stevens, S. S. (1975). Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects. Piscataway, NJ: Transaction Books. ISBN: 978-0887386435

Summers, I. (Ed.). (1996). Tactile aids for the hearing impaired. Whurr. ISBN: 978-1-870332-17-0

Tan, H. Z., & Rabinowitz, W. (1996). A new multi-finger tactual display. Journal of the Acoustical Society of America, 99(4), 2477–2500. https://doi.org/10.1121/1.415560

Tan, H., et al. (2003). Temporal masking of multidimensional tactual stimuli. *Journal of the Acoustical Society of America, 116*(9), 3295–3308. https://doi.org/10.1121/1.1623788

Tan, H., Rabinowitz, W., & Verrillo, R. T. (1992). Perception via the sense of touch. In Summers, I. R. (Ed.), Tactile Aids for the Hearing Impaired (pp. 1–36). Whurr. ISBN: 978-1-870332-17-0

Vander Poorten, E. E. B., Demeester, E., & Lammertse, P. (2012). Haptic feedback for medical applications: A survey. In Actuator Conference. Bremen, Germany. https://www.radhar.eu/publications/e.-vander-poorten-actuator12-haptic-feedback-for-medical-applications-a-survey

Vasarhelyi, G., et al. (2006). Effects of the elastic cover on tactile sensor arrays. Sensors and Actuators A: Physical, 132(1), 245–251. https://doi.org/10.1016/j.sna.2006.01.009

Wagner, C., Stylopoulos, N., & Howe, R. (2002). The role of force feedback in surgery: Analysis of blunt dissection. In 10th Symposium on Haptic Interfaces for Virtual Environment and

Teleoperator Systems (pp. 53–60). https://doi.org/10.1109/HAPTIC.2002.998943

Wagner, C., Stylopoulos, N., & Howe, R. (2002). The role of force feedback in surgery: Analysis of blunt dissection. In 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (pp. 53–60). https://doi.org/10.1109/HAPTIC.2002.998943

Westebring-van der Putten, E. P., et al. (2008). Haptics in minimally invasive surgery - a review. Minimally Invasive Therapy & Allied Technologies, 17(1), 3–16. https://doi.org/10.1080/13645700701820242

Wolf, U. (2007). Aristoteles' Nikomachische Ethik. Darmstadt: Wissenschaftliche Buchgesellschaft. ISBN: 978-3-534-20824-2

Grunwald, M. (2008). Human haptic perception: Basics and applications. Basel: Birkhäuser. ISBN: 3764376112

Kant, I. (1983). Anthropologie in pragmatischer Hinsicht: Philosophische Anthropologie; Charakter. Stuttgart: Reclam. ISBN: 3-15-007541-6

Getzinger, G. (2006). Haptik: Rekonstruktion eines Verlustes. Wien: Profil. ISBN: 3-89019-521-0

Néret, G. (1998). Erotik in der Kunst des 20. Jahrhunderts. Tokyo: Taschen. ISBN: 3-8228-7853-7

Young, S. N., & Peschel, J. M. (2020). Review of human-machine interfaces for small unmanned systems with robotic manipulators. IEEE Transactions on Human-Machine Systems, 50(2), 131–143. https://doi.org/10.1109/THMS.2020.2969380

CHAPTERVII

The Importance of Choosing the Right Features for Machine Learning Models in a Business

Hasan Tahsin AYTEKİN[1]

1. INTRODUCTION

Machine learning has gained an important place in the decision-making processes of enterprises and has become an indispensable tool in terms of technology management. In this context, one of the most critical steps that determine the success of machine learning models is feature selection. Feature selection enhances the competitive advantage of enterprises by optimizing the performance of models. In this study, the effects of feature selection on technology management in enterprises and how this process can be integrated into strategic decision-making mechanisms will be examined.

As the size of the data increases, the application of feature selection methods is the most important method for reducing the size (Parlak & Uysal, 2021). The purposes of feature selection are to

[1] Ufuk University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Orcid: 0000-0002-5632-7825, hasan.aytekin@ufuk.edu.tr

develop simpler and more interpretable models, enhance data mining performance, and ensure the data is clean and comprehensible (Li, et al., 2017). Feature selection involves the identification of accurate and relevant data in the process of developing machine learning models. This increases the prediction accuracy of the models and enables businesses to make more accurate decisions. Especially when working with large datasets, the selection of the right features helps models avoid overfitting and increase their generalization ability. Moreover, by using fewer features, it is possible to increase the interpretability of the models and their understandability by management.

Within the scope of technology management strategies, feature selection is also of great importance in terms of cost efficiency and efficient use of resources. The data structure simplified by feature selection prevents overfitting and thus, the need for optimized reduced data collection and storage of collected data decreases computational, processing and operational costs, resulting in cost savings in enterprises (Jiang, Che, He, & Yuan, 2023).

Another critical aspect of feature selection is to ensure compliance with business goals and strategic objectives. Determining the features that are compatible with the strategic goals of businesses enables the development of more effective and targeted machine learning models. In addition, in terms of data privacy and regulatory compliance, using only the necessary data helps businesses to comply with legal requirements.

Finally, feature selection also offers a great advantage in terms of adaptability and scalability of machine learning models. Models that can quickly adapt to changing business conditions and can be easily scaled across different units provide flexibility to businesses. This study aims to address all these aspects of feature selection and to demonstrate how it can add value to businesses in the context of technology management.

1.1 Definition of the Problem

Feature selection is one of the most important steps in the development of machine learning models. Many methods developed for feature selection are continuously improved and new ones are added to these methods with the emergence of new needs of rapidly growing and changing datasets (Ali, et al., 2024). For the evaluation of feature selection algorithms, the stability of the models is an important factor as well as the importance of model quality (Buyukkececi & Okur, 2023).

The selection of features is one of the most significant factors influencing the efficacy of artificial intelligence (AI) models employed to enhance technology management. Technology management entails the integration of knowledge and expertise to inform decisions regarding technology strategy, innovation, and management, with the objective of securing a sustainable competitive advantage. It is essential to evaluate the impact of feature selection in machine learning models on enterprise operations, the value it offers, and its influence on technology management costs. This evaluation should be conducted using a specific feature selection application and the results should be interpreted.

1.2 Purpose of the Study

The purpose of investigating the importance of choosing the right features for machine learning models in a business is to enhance the efficacy of machine learning applications in business contexts. This is done in order to optimize business processes, support strategic decision-making, ensure resource efficiency, guarantee compliance with regulations, achieve competitive advantages and contribute to academic knowledge. This study seeks to bridge the gap between technical advances and strategic business needs.

2. CONCEPTUAL FRAMEWORK

The theoretical foundations of the study on the importance of choosing the right features for machine learning models in a business consists of several interdisciplinary fields, including machine learning theory, management science principles, technology management, data science, economic theory, behavioral science, and ethical and regulatory frameworks. These fields provide a comprehensive basis for demonstrating how feature selection can improve business processes, decision making, innovation, resource allocation and compliance, thus bridging the gap between technical advances and strategic business needs.

2.1 Machine Learning Theory

Numerous recent articles consider machine learning to be synonymous with artificial intelligence. However, artificial intelligence is a broad term that encompasses everything from intricate rule-based software to human-level intelligence that has yet to be developed. In reality, machine learning is a specific subset of AI focused on developing programs that learn from data rather than relying on predefined rules, enabling software to learn from examples. Its algorithms aren't explicitly programmed like traditional software; instead, they are trained on extensive sets of relevant data. (Aytekin, 2021).

2.1.1 Bias-Variance Tradeoff

When a problem needs machine learning discipline to be utilized for a feasible solution, the distribution properties of the dataset for the problem may not fit to a specific pattern. In order to minimize the errors generated at the testing phase of model development, optimization of the model itself is a common practice. The goal of model optimization for the purpose of reducing test errors can result in an increase in the frequency of incorrect predictions during the training phase of the model. This may lead to the emergence of overfitting, a phenomenon characterized by the model's inability to generalize effectively. Consequently, the selection of a model based only on its test performance may not necessarily guarantee its generalizability (Guan & Burton, 2022).

The bias-variance trade-off indicates that a model must strike a balance between underfitting and overfitting: It expresses the underlying structure in the data and avoids fitting spurious patterns in the data (Belkin, Hsu, Ma, & Mandal, 2019).

It is very important to understand how feature selection affects the trade-off between bias and variance. Reducing irrelevant features reduces model complexity and variance, leading to better generalization.

2.1.2 Dimensionality Reduction

Dimensionality reduction is a preprocessing technique in machine learning that eliminates redundant features, as well as noisy and irrelevant data, to enhance feature learning accuracy and reduce training time. To apply dimensionality reduction techniques, methods such as feature selection and feature extraction are utilized (Velliangiri, Alagumuthukrishnan, & Joseph, 2019).

Theories about techniques such as Principal Component Analysis (PCA) and Lasso regression are crucial in understanding how to efficiently minimize the number of features while retaining the critical information.

2.1.3 Model Interpretability

The concepts of Explainable Artificial Intelligence (XAI) emphasize the importance of building models that are not only accurate but also interpretable by non-technical stakeholders, which is directly influenced by feature selection.

In order to make neural network and deep learning models interpretable, the stages of rule extraction and then model exhibition are generally applied (Huang, et al., 2020). The extraction process is about discovering the relevant information and making it usable for the next step. The exhibition process is the organization of the information that is discovered and organized in a way that can be used for the targeted process in a way that people can understand (Gao & Guan, 2023).

2.2 Principles of Management Science

2.2.1 Decision Theory

Each decision has a variable impact on the organization in question. Transparency is very important to increase the impact of the decisions taken (Tsirtsis, et al., 2024). In real life, decision making is often carried out by comparing different options against various criteria. In many cases, the criteria used for comparison may also contradict each other. This situation increases the uncertainty in the decision-making process. Due to the increased uncertainty, decision makers are required to conduct comprehensive research on the significance of conflicting criteria (Dorini, Kapelan, & Azapagic, 2011).

Individuals meet their needs in the most appropriate way they think with the rational choices they make. Utility is fundamentally a psychological concept and serves as a foundational element in economics and finance. Utility theory explains the rational choices of individuals (Akkaya, 2021). The explanation put forward by utility theory is expressed using the unit type called 'utility'. In this context, 'utility' is the value obtained from the measurement of the level of satisfaction from consuming the relevant good or service.

On the other hand, risk factors also play an important role in managers' decision-making processes. Enterprise risk management aims to improve decision-making processes and help organizations avoid malignant problems (Crawford & Jabbour, 2024). Theories of decision-making under uncertainty, utility theory, and risk management provide a framework for understanding how feature selection can improve managerial decision-making processes.

2.2.2 Operational Efficiency

Operations research principles are applied to analyze how feature selection can optimize processes and resource allocation and increase overall efficiency. The recent developments in the field of operations research include data science, machine learning and artificial intelligence. These areas of development have considerable potential for the future and are anticipated to change various aspects of the operations research domain. Machine learning-based heuristics are already being extensively used to address various complex problems. (Petropoulos, et al., 2024). Artificial intelligence is transforming all kinds of industries by significantly reducing costs and increasing operational efficiency (Prabhod, 2024).

2.2.3 Strategic Management

The primary branch of strategic management is strategic planning. The principles of strategic management determine the guidance, direction and boundaries of operational management. The field of strategic management is concerned with the formulation of operational strategies, whereas strategic planning is the process of implementing operational decisions. However, as in strategic management, the focus and emphasis of strategic planning is strategy rather than operations (Steiner, 1979). The critical features that define the strategic nature of strategic planning should be seen as conditions to be targeted. These include the content, the process's structure, its relational aspects, and the institutional consequences (Albrechts & Balducci, 2013). Theories on strategic planning and competitive advantage emphasize the role of accurate data-based insights in formulating and executing business strategies.

2.3 Technology Management

2.3.1 Innovation Management

Theories about technological innovation and its diffusion within the organization provide a framework for how advanced machine learning techniques facilitated by feature selection can foster innovation in organizations. Innovation is now happening faster than ever before and companies are forced to accelerate their innovation efforts to maintain their competitive advantage and open up to new products, services, and markets (Bellis, Magnusson, Nilsson, & Samuelsson, 2024). One of the driving forces that accelerates the pace of innovation in many sectors is generative artificial intelligence and the pace of change in generative artificial intelligence is also accelerating. Generative artificial intelligence not only creates new use cases for businesses but can also accelerate or scale existing ones (McCausland, 2024).

2.3.2 Technology Adoption

Information technologies provide increased performance in all areas, both for individual and corporate use. The technology acceptance model (TAM) provides a comprehensive explanation of the various factors that influence user acceptance of information systems. TAM assumes that the perceived usefulness of a technology in performing a task is determined by its ease of use and predicts the acceptance of the technology by users on the basis of their behavioral intentions. The main purpose of TAM is to provide a theoretical explanation for understanding the successful implementation of technology. Additionally, it seeks to anticipate user behavior by elucidating the underlying mechanisms that influence the acceptance of technology. The goal of TAM is to inform implementers about the precautions they can take before the implementation of systems. The TAM represents perceived ease of use, perceived usefulness and behavioral intention, which represent individuals predicted behavior (Davis, 1989). Models such as the Technology Acceptance Model (TAM) are used to examine how feature selection affects the adoption and use of machine learning technologies in business contexts.

2.4 Data Science and Information Systems

2.4.1 Data Quality

The field of data science merges knowledge from statistical, mathematical, computer and behavioral science to derive insights from organizational data. Data science tools used to predict possible future outcomes are called predictive analytics. (Hazen, Boone, Ezell, & Jones-Farmer, 2014).

The term data quality is also applicable to the entity of data itself. It determines the accuracy, consistency, timeliness, correctness, and relevance of the data within its specific context. While the significance of data quality may appear self-evident, its practical implementation is not an easy task. The rapid increase in the number of data sources and the accelerated rate of data production highlight the essential importance of data quality in this context.

It is essential that users place their trust in the data produced by information systems that integrate software applications to support the processes used within their organizations. In order to ensure that the data they produce and manage within their organizations' information systems is of an appropriate quality level. Therefore, it is a fundamental requirement for every organization that the data to be managed by information systems must be of sufficient quality. In order to ensure a minimum level of data quality in software development projects, it is essential to define rules and properties governing the creation, utilization and storage of data assets. This is the only means of guaranteeing the quality of data assets managed by the information system. The concept of data quality is highly dependent on the context in which data is produced or used and is inherently multidimensional (Guerra-García, et al., 2023). Theories concerning data quality, including completeness, accuracy, and relevance, underscore the significance of feature selection in delivering high-quality input data for machine learning models.

2.4.2 Information Value

The importance of uncertainty for a decision maker is determined not only by the probabilistic nature of the uncertainties, but also by the economic impact that these uncertainties will create (Howard, 1966). The value of knowledge arises as a result of choices made in uncertain circumstances. When individuals believe that their significant assets are exposed to an uncertain degree of risk, they are motivated to seek out additional information to eliminate that risk. If the gain from additional information is greater than the cost of obtaining the information, they are always willing to pay for it. When the general results obtained from information models on the value of information are analyzed, it is seen that the value of information largely depends on factors related to how uncertain decision makers are about the value of information, what is at stake as a result of their decisions, how much it costs to use information to make a decision, and the price of another information that is considered to be the most accurate that can be used instead of certain information (Macauley, 2005).

Information value analysis is a data exploration technique that helps to identify which columns in a given dataset are capable of predicting the value of a specific dependent variable. Concepts used in information theory, such as the value of information, help to understand how selecting the most informative features can improve decision making and business outcomes.

2.5 Economic Theories

Economic principles are used to evaluate the trade-offs involved in feature selection, weighing the costs of data collection and processing against the benefits of improved model performance and decision making.

2.5.1 Cost-Benefit Analysis

The purpose of cost-benefit analysis is to establish a consistent framework of decision rules for assessing choices based on their outcomes and to examine the consequences of a decision obtained through these rules in terms of cost-benefit (Drèze & Stern, 1987). In practice, it is the process of weighing overheads against earnings where all relevant costs are formulated and collected. Then all the benefits to be derived from the execution of a particular project are tabulated and summed up. The decision to execute the project is taken only when the gains from the execution of the project are greater than the costs that will arise from the implementation of the project (Odek & Oluoch, 2023).

2.5.2 Resource Allocation

Resource allocation is the allocation of limited resources to specific purposes selected from various applicable possibilities. Since resources are limited, it is very important to choose which goods and services to produce in order to ensure efficiency.

Resource-based views and efficient resource allocation theories provide a framework for understanding how feature selection helps optimize the use of organizational resources. The adoption, adaptation and maintenance of technology in the context of resource allocation is a key focus for organizations. Organization's main objective with the resource allocation on technology is to facilitate the informed decision-making process regarding these aspects. Considering the long-term technology-focused strategic plans of organizations, it is possible that the costs associated with making urgent resource allocations in the short term may exceed those initially anticipated. Organizations that use digital technologies effectively want to optimize resource allocation by using the most efficient decision-making processes they can create for the purpose of maximizing their competitive advantage (Porath, 2023). The conventional managerial approaches employed by traditional management teams in organizations are being reshaped due to digital transformation. Digital technologies, which are among the factors that make digital transformation possible, produce a vast quantity of digital data originating from both the human element and the hardware itself. The goal of facilitating the interpretation of this data produced in organizations and ensuring that the large amount of information obtained can be used by management staff, necessitates a transition to data-driven decision making. Managers who utilize data-driven decision-making mechanisms will be able to optimally distribute scarce resources by employing advanced analytical techniques and artificial intelligence applications.

2.6 Behavioral Sciences

2.6.1 Cognitive Load Theory

Cognitive load theory is a theory of instruction based on our knowledge of human cognition. Since its emergence in the 1980s, it has been using human cognitive architecture to create experimental and instructional effects (Sweller J. , 2011). Cognitive load theory (Sweller J. , 2004) hypothesizes that working memory is too limited to cope with new information and long-term memory is unlimited to hold complex cognitive schemas. The capacity of working memory is limited during the storage of information, whereas long-term memory is not. The knowledge stored in cognitive schemas is much more important than the numerous new items in long-term memory that are not organized. Human reasoning ability is derived from the knowledge stored in cognitive schemas, rather than from new items in long-term memory (Van Merrienboer & Ayres, 2005). When the cognitive load theory is analyzed from the perspective of instructional psychology, it is seen that the knowledge obtained about the processing limitations of the human cognitive system is used to improve the effectiveness of instructional procedures and the presentation of knowledge. The assumptions and principles that support cognitive load theory are founded upon an evolutionary perspective that regards both the human cognitive structure and biological evolution as natural information processing systems (Kalyuga, 2011).

This theory explains how reducing the number of features simplifies models, making it easier for managers to understand them and use them effectively in decision making. The reduction of features used in machine learning models is a complex task. The complexity of a task affects every living being in terms of the performance achieved as a result of the behavior performed to accomplish the task. This performance is measured in terms of the number of interactive elements in learning materials presented in cognitive load theory. An element is defined as any entity that requires processing and learning. It is not possible to process and learn about interacting elements in isolation. This is the most decisive behavioral feature of interaction. All interacting elements must be processed simultaneously in order for them to make sense. In accordance with the cognitive load theory, the cognitive load is determined by the number of elements that must be processed concurrently (Chen, Paas, & Sweller, 2023).

2.6.2 Behavioral Decision Making

Insights from behavioral economics and psychology are applied to examine how feature selection affects decision makers' cognitive processes, their trust and confidence in machine learning models. The decision-making process represents a fundamental cognitive function which is an integral component of human behavior. It involves the selection of a preferred option or course of action from a range of alternatives based on the application of specific criteria. Decision theories have been widely used across various disciplines, including computer science, cognitive informatics, economics, management science, psychology, sociology, statistics and political science (Wang & Ruhe, 2007). At the same time, it constitutes a prominent cognitive process. Among many features of the cognitive process, the selection process and the basic cognitive processes are the most distinctive ones. The basic cognitive processes may be conceived of as memory, perception and attention. The selection itself as a cognitive process like perception or judgement. However, the process of selecting among alternatives starts with the action of selection and ends with another action, which is contingent upon the selection made. In order to select an action among many potential alternatives, it is first necessary to evaluate the alternative values and then to identify the most appropriate and feasible value. The consequence of any selection is either a return of feedback or the acquisition of knowledge from the result of that selection. (Gonzalez, 2014).

Theories that provide a detailed explanation of the psychological processes underlying human decision-making process are referred to as behavioral decision theories. Although it is referred to as a theory, it is a blend of multiple psychological theories that are generally limited to qualitative information (Takemura, 2014). Behavioral decision making is the study of the cognitive, emotional, and social processes used in identifying and selecting among alternatives. In order to make a definitive choice, the decision maker directs the decision-making processes in line with his/her preferences.

2.7 Ethical and Regulatory Frameworks

2.7.1 Data Privacy and Ethics

The term "Artificial Intelligence" (AI) is used to describe the capacity of a computer to undertake tasks that would typically be carried out by humans. These tasks include problem-solving, planning, reasoning and learning from experience (Wang P. , 2019). The growing deployment of AI technologies in organizations is also influencing how and why people engage with their work, as well as the perceived meaningfulness of their work. However, despite the increasing prevalence of AI in organizations, there is no established understanding or set of rules that governs how the use of AI will affect work environments and the ethical implications of the changes in business practices brought about by the use of AI (Bankins & Formosa, 2023).

The theoretical frameworks surrounding data privacy, ethical AI, and legal compliance provide a foundation for understanding the implications of feature selection on data use and privacy concerns. The United Nations has confirmed in Article 12 of the Universal Declaration of Human Rights that privacy continues to be a human right in the digital age. However, our daily digital experiences and the seemingly ever-increasing amounts of data show that privacy is a concept that is commonplace, distributed, and violated through technology (Gstrein & Beaulieu, 2022).

The relationships between the collective and the individual, the group and the individual, and the public and the private are increasingly driven by the production and commodification of data. Globally, many different scientific, societal and regulatory frameworks are being used to manage and change these processes. For example, effective privacy laws or data protection regulations have been developed, such as the European Union's General Data Protection Regulation, which regulates the principles of collection of information by whom, to what extent and for what purposesD (Gstrein & Beaulieu, 2022).

On the other hand, when data processing is intended for scientific research, the obligation to obtain the consent of individuals for the processing of personal data plays a very important role, due to the characteristics of the General Data Protection Regulation of the European Union applicable in the research context. Nevertheless, the information obligations introduced in the General Data Protection Regulation of the European Union are not entirely satisfactory and contain some flaws (Ducato, 2020).

The European Union's General Data Protection Regulation, which was developed due to the increasing use of data consisting of personal information about individuals, has led many countries, and thus, as of early 2022, 157 countries have developed legislation to protect personal data and the privacy of this data (Demirer, Hernández, Li, & Peng, 2024).

2.7.2 Compliance Theories

Ensuring that feature selection in machine learning models complies with legal standards and ethical guidelines is a critical aspect of developing responsible and compliant AI systems.

2.7.2.1 Regulatory Requirements

Data Protection Laws: Regulations like the General Data Protection Regulation (GDPR, 2018) in the EU and the California

Consumer Privacy Act (CCPA, 2018) in the US set strict rules on the collection, processing and storage of personal data. Feature selection must comply with these laws by ensuring that sensitive personal data is adequately protected and used only for legitimate purposes.

Fairness and Non-Discrimination: Laws such as the Equal Credit Opportunity Act (ECOA, 1974) in the US prohibit discrimination based on protected characteristics. Feature selection processes must avoid using features that may lead to biased results or indirect discrimination.

Transparency and Disclosability: Regulatory frameworks (CCPA, 2018) (ECOA, 1974) (GDPR, 2018) generally require that decisions made by AI systems be explainable. This requirement means that the features used in the model should be interpretable and the logic behind their selection should be clear to stakeholders.

Accountability and Governance: Organizations are required to implement governance frameworks that ensure accountability for AI systems, such as ISO/IEC 42001:2023 information technology AI management system (ISO/IEC 42001:2023, 2023), ISO/IEC 23894:2023 guidance on risk management for information technology AI (ISO/IEC 23894:2023, 2023) and ISO/IEC 23053:2022 framework for AI systems using machine learning (ISO/IEC 23053:2022, 2022). This includes documenting feature selection processes, conducting regular audits, and maintaining compliance with internal and external standards.

2.7.2.2 Compliance Theories

Risk-Based Approach: The risk-based approach to compliance allows businesses and organizations to assess their potential to violate laws, regulations, and policies, ensuring that resources are directed to the areas where they are needed most and will be most effective. Risks of non-compliance are identified, and then appropriate compliance measures are implemented to control these risks. This theory suggests that organizations should assess and mitigate the risks associated with feature selection. This includes identifying potential ethical and legal risks, assessing their impacts and implementing measures to minimize these risks.

Value-Sensitive Design: Value-Sensitive Design is a theoretically based approach that represents a pioneering effort to systematically and principledly consider human values throughout the technology design process (Friedman, Hendry, & Borning, 2017). This approach emphasizes the incorporation of human values into the design and implementation of technology. Values such as fairness, privacy, and transparency should be considered in feature selection and the resulting models should be compliant with societal and ethical standards.

Ethical AI Frameworks: Many institutions and organizations are developing ethical AI frameworks to guide them in developing responsible AI projects. The frameworks developed are evaluated through accountability, responsibility, and culpability (O'Sullivan, et al., 2018). These frameworks usually include principles such as usefulness, nonmaleficence, autonomy, and fairness that should be considered during feature selection.

Stakeholder Engagement: Organizations do not have the option of not engaging with stakeholders, and the only thing they can control is when and how successfully they engage (Jeffery, 2009). The process of engaging relevant stakeholders of an organization with a well-defined purpose inorder to achieve shared goals is called stakeholder engagement. (Leopizzi, 2023). Engaging with stakeholders (e.g., customers, employees, regulators) throughout the feature selection process can help ensure that the selected features align with societal expectations and regulatory requirements. This includes seeking feedback and incorporating different perspectives.

2.8 Dimension and Curse of Dimensionality

The concept of dimensionality mentioned in this study refers to the attributes or features in the datasets used for machine learning purposes. The curse of dimensionality is a concept which states that when the quantity of data used to train a model is fixed, increasing the dimensionality can result in overfitting, which in turn can reduce classification success rates (Anuragi, Sisodia, & Pachori, 2024). This also means that patterns and relationships that are valid in a restricted domain may be invalid in a complex domain.

3. METHOD

3.1 Research Model and Hypotheses

The aim of this research is to empirically demonstrate the importance of feature selection used in machine learning models. Feature selection can improve the performance of machine learning models, reduce overfitting, and increase the generalization ability of the model. The following hypotheses will be tested in the research:

- **H1:** Feature selection improves the accuracy of machine learning models.
- **H2:** Different feature selection methods provide different performance improvements in different machine learning models.
- **H3:** Reducing the number of features prevents overfitting of the model and increases the generalization ability.

The research model is based on comparing the performances of feature selection methods and various machine learning classifiers. This comparison will be based on accuracy rates.

3.2 Universe and Sample

The universe of the research consists of datasets used to examine the performance of feature selection and classification methods in machine learning models. A dataset obtained from the ‘UCI Machine Learning Repository’ platform will be used as a sample (Kelly, Longjohn, & Nottingham, 2024). The following criteria will be taken into consideration in the selection of the dataset:

- **Data Diversity:** The dataset should exhibit a range of features and comprise a sufficient number of samples.
- **Data Quality:** Minimization of missing or erroneous data.
- **Problem Type:** The dataset should contain a classification problem.

The dataset selected for illustrative purposes is the 'Breast Cancer Wisconsin (Diagnostic)' dataset, which is located in the 'UCI Machine Learning Repository' environment. This is a widely used dataset in the field of health, specifically in the context of breast cancer research (Wolberg, Mangasarian, Street, & Street, 1995)

(Street, Wolberg, & Mangasarian, 1993). The analyses to be performed on this dataset will help test the hypotheses of the research.

3.3 Data Collection Methods and Tools

The ‘Breast Cancer Wisconsin (Diagnostic)’ dataset will be downloaded from the Kaggle platform (Kaggle, 1995), which is used as an external data sharing environment by the ‘UCI Machine Learning Repository’ and analyzed using the Python programming language. The main libraries and tools to be used are:

- **scikit-learn:** Will be used for machine learning models and feature selection.
- **pandas:** Will be used for data manipulation and analysis.
- **numpy:** Will be used for numerical operations.
- **matplotlib:** Will be used for data visualization.

The data collection process will include the following steps:

1. **Downloading the Dataset:** The appropriate dataset will be downloaded from the Kaggle platform.
2. **Data Preprocessing:** The dataset will be cleaned, missing data will be filled or removed, and the data will be normalized.
3. **Feature Selection:** The objective is to identify a subset of features that optimally represent the dataset. The optimal number of features will be selected using the SelectKBest class in the scikit-learn library. Given that the dataset comprises 31 features, the initial stage will entail the selection of the single best feature. This will then be followed by the incremental selection of the two best features, the three best features, and so forth, until the 31 best features have been identified. The results will then be tabulated for analysis, with the cross-validation scores used to assess the accuracy of the predictions and the model performances with different numbers of selected features.

3.4 Data Analysis

The analysis of the data will be performed by feature selection and comparison of the performance of various machine learning classifiers. This process includes the following steps:

1. **Feature Selection:** The optimal number of features will be identified through the utilization of the f_classif score function within the SelectKBest class.
2. **Model Training:** Using the selected features, machine learning classifiers will be trained using K-Nearest Neighbors, Gaussian NB, Neural Network, SVM-SVC and SVM-NuSVC models.
3. **Model Evaluation:** The performance of the models will be evaluated with accuracy rates and cross-validation method.
4. **Comparison of Results:** The obtained accuracy rates will be compared and analyzed according to the number of features selected and the classifier methods used. For this purpose, the data will be divided into training and validation sets, the relevant model will be trained with training set and the expected performance of the model on new data will be estimated by evaluating the validation set. In this study, cross-validation will be applied for each feature set and the average accuracy rate of each model will be calculated. These accuracy rates will be visualized with tables and graphs according to the number of features and the classifier used.

The objective of this method section is to enhance the reliability of the findings by guaranteeing that the research is conducted in a systematic and comprehensive manner. The research endeavors to substantiate the pivotal function of feature selection on the performance of machine learning models with experimental data.

4. FINDINGS

Two Python classes were constructed for the purpose of conducting this research and evaluating the resulting data. The prepared Python classes will be used to perform feature selection, and the accuracy rates of machine learning classifier models (K-

Nearest Neighbors, Gaussian NB, Neural Network, SVM-SVC and SVM-NuSVC) will be compared using different numbers of selected features.

The SelectKBest method, which is part of the scikit-learn.feature_selection library, is a feature selection method that identifies the k best features based on a scoring function within the scikit-learn library. In this study, the scoring function employed is f_classif. The SelectKBest method enables the removal of irrelevant or superfluous features, thereby reducing the dimensionality of the dataset and enhancing the performance of the model.

The Python environment software utilized to generate the comparison data for feature selection is detailed in Section 4.2.

4.1 Hypotheses to be Evaluated

The following hypotheses will be tested in the research:

- H1: Feature selection improves the accuracy of machine learning models.
- H2: Different feature selection methods provide different performance improvements in different machine learning models.
- H3: Reducing the number of features prevents overfitting of the model and increases its generalization ability.

4.2 Feature Usage Comparison Data Generation Software

The details of the program created to compare the accuracy rates of machine learning classifiers K-Nearest Neighbors, Gaussian NB, Neural Network, SVM-SVC and SVM-NuSVC models using different numbers of features are given in the following subsections.

4.2.1 Loading Necessary Libraries

The program is created from software codes implemented in the Python environment. The libraries that need to be loaded in order for all program parts to work are given below.

```
# Required libraries in the main section
import pandas as pd
import matplotlib.pyplot as plt


# Libraries required in the dataset loading
class (LoadData)
import pandas as pd
from sklearn.preprocessing import
StandardScaler


# Libraries required in the evaluation class
(EvaluateModels)
import pandas as pd
from sklearn import svm
from sklearn.metrics import accuracy_score
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import
KNeighborsClassifier
from sklearn.linear_model import Perceptron
from sklearn.model_selection import
train_test_split, cross_val_score
from sklearn.feature_selection import
SelectKBest, f_classif
```

4.2.2 Downloading and Preprocessing the Dataset

After the dataset is downloaded from the Kaggle platform (Kaggle, 1995), which is used as an external data sharing environment by the ‘UCI Machine Learning Repository’ described in Section 3.3, it should be added to the Python project environment. The raw data is named ‘BreastCancer.csv’ in the codes given in this paper. The preliminary preparations required to process the data are performed in the Python class called ‘LoadData’ as in the software codes given below. At this stage, the errors in the data will be fixed and loaded into the Pandas data frame. In order for the loaded data to be used in the operations, its identifier states must be loaded into the class instance variables as performed in the LoadData class below.

```
class LoadData:
    def __init__(self, file_name,
decision_feature):
        self.x_col_names = None
        self.X = None
        self.y = None
        self.fileName = file_name
        self.process_normalized_data(
                            decision_feature)

    def split_column_names(self, df,
first_list_columns):
        # Splits the column names in the data
frame into two lists based on the given column
names.
        all_columns = df.columns.tolist()
        first_list = [col for col in
first_list_columns if col in all_columns]
        second_list = [col for col in
all_columns if col not in first_list]

        return first_list, second_list

    def process_normalized_data(self,
decision_feature):
        # Loading the dataset
        data = pd.read_csv(self.fileName)

        # Examining the dataset
        print(data.head())

        # Data preprocessing
        X = data.iloc[:, 0:data.shape[1]]
        y = data.iloc[:, -1]

        # Normalize data
```

```
        # Assign values to class instance
variables
        self.x_col_names = X.columns

        decision_feature_list,
list_wo_decision_feature =
self.split_column_names(X, [decision_feature])

        self.X = X[list_wo_decision_feature]
        self.y = X[[decision_feature]]

    def get_data(self) -> object:
        return self.X, self.y
```

4.2.3 Feature Selection and Model Training

After the dataset to be used is ready in the Python environment, the first thing to do is to select the features and then train various machine learning classifiers and record their accuracy rates. The EvaluateModel class, whose Python code is given below, stores the results of the cross-validation rates in order to evaluate the training and application results of each classifier model so that they can be used in visualization and evaluation. The details of these steps to be taken for the preparation of the analysis data are given in Section 3.4.

```
from LoadData import LoadData

class EvaluateModels(LoadData):
    def __init__(self, data_file_name,
decision_feature):
        LoadData.__init__(self, data_file_name,
decision_feature)
        self.Column_Names_KBest = None
        self.dataFileName = data_file_name

    def print_SelectKBest_scores_all(self,
number_of_features):
        if number_of_features == 0:
```

```
number_of_features = self.X.shape[1]

        # Determining the features to be used
in feature selection
        selector =
SelectKBest(score_func=f_classif,
k=number_of_features)
        selector.fit(self.X,
self.y.values.reshape(-1,))

        # Score values of processed features
and their order in ascending order
        dfscores =
pd.DataFrame(selector.scores_)
        dfcolumns =
pd.DataFrame(self.x_col_names)
        feature_scores = pd.concat([dfcolumns,
dfscores], axis=1)
        feature_scores.columns = ['Feature',
'Score']

        self.Column_Names =
self.X.columns.tolist()

        print('')
        print('Feature scores (top ' +
str(number_of_features) + ')')
        print('')
        print(feature_scores.nlargest(
number_of_features, 'Score',"all"))
        print('')

print('===============================================
====================')
        print('')

        feature_scores.to_csv('Feature Scores
+' + str(number_of_features) + ".csv", sep=";",
decimal=',', index=True, index_label="Feature")
        pass
```

```
    # Feature selection and model training
method
    def evaluate_models(self,
number_of_features):
        # Feature selection
        selector =
SelectKBest(score_func=f_classif,
k=number_of_features)
        X_new = selector.fit_transform(self.X,
self.y.values.reshape(-1,))

        # Identification of the models
        models = {
            'K-Nearest Neighbors':
KNeighborsClassifier(n_neighbors=5),
            'Gaussian NB': GaussianNB(),
            'Neural Network':
Perceptron(random_state=1),
            'SVM-SVC': svm.SVC(kernel="rbf",
C=0.025, probability=True),
            'SVM-NuSVC':
svm.NuSVC(probability=True)
        }

        results = {}

        # Training and evaluation of the models
        for model_name, model in
models.items():
            print(model_name + " [" +
str(number_of_features) + "]")
            scores = cross_val_score(model,
X_new, self.y.values.reshape(-1,), cv=10,
scoring='accuracy')
            results[model_name] = scores.mean()

        return results
```

4.2.4 Main Program Flow

After creation of the processing classes in the Python environment, the software codes required to load the process data into the system and then perform the tests with different numbers of features and visualize the results are as follows:

```
from EvaluateModels import EvaluateModels

evaluator = EvaluateModels('BreastCancer.csv',
'diagnosis')

evaluator.print_SelectKBest_scores_all(0)

# Evaluation of models with different number of
features
number_of_features_list = [1, 2, 3, 4, 5, 6, 7,
8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19,
20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31]


all_results = {}


# Evaluate all models with different number of
features
for number_of_features in
number_of_features_list:
    all_results[number_of_features] =
evaluator.evaluate_models(number_of_features)

# Visualization of results
results_df = pd.DataFrame(all_results).T
results_df.plot(kind='bar')
plt.xlabel('Seçilen Özelliklerin Sayısı')
plt.ylabel('Doğruluk')
plt.title('Farklı Sayıda Seçili Özelliklerle
Model Performansı')
plt.legend(loc='lower right')
plt.show()
```

```
      # Storing results
print(results_df)
results_df.to_csv('Model Performance +' + str(number_of_features_list[-1]) + ".csv", sep=";", decimal=',', index=True, index_label="Features")
```

4.3 Comparison Data

Within the framework of the model and hypotheses of the research given in Section 3.1, the prediction accuracy rates of the models were obtained by using different number of features of different machine learning models with the 'Breast Cancer Wisconsin (Diagnostic)' dataset containing real life data specified in Section 3.2 using the software described in Section 4.2.

4.3.1 Feature Rating Scores

Table 1 lists the scoring values of various features. Using the features in this table, the results of the accuracy rates obtained according to the number of features used in the training of machine learning classifiers K-Nearest Neighbors, Gaussian NB, Neural Network, SVM-SVC and SVM-NuSVC algorithms will be generated. The table shows that 'concavity_worst' and 'texture_worst' features have the highest scores, while 'fractal_dimension_worst' and 'concave_points_se' features have the lowest scores.

Table 1. Feature rating scores

No	Feature	Score	No	Feature	Score
28	concavity_worst	964,39	25	area_worst	122,47
23	texture_worst	897,94	29	concave_points_worst	118,86
8	concavity_mean	861,68	2	radius_mean	118,10
21	fractal_dimension_se	860,78	18	concavity_se	113,26
3	texture_mean	697,24	5	area_mean	83,65
24	perimeter_worst	661,60	9	concave_points_mean	69,53
1	diagnosis	646,98	30	symmetry_worst	66,44
4	perimeter_mean	573,06	16	smoothness_se	53,25

7	compactness_mean	533,79	17	compactness_se	39,01
27	compactness_worst	436,69	20	symmetry_se	3,47
6	smoothness_mean	313,23	15	area_se	2,56
26	smoothness_worst	304,34	0	id	0,90
11	fractal_dimension_mean	268,84	10	symmetry_mean	0,09
13	texture_se	253,90	12	radius_se	0,04
14	perimeter_se	243,65	19	concave_points_se	0,02
22	radius_worst	149,60	31	fractal_dimension_worst	

Source: Table created by the author

4.3.2 K-Nearest Neighbors Model Accuracy Rates

Table 2 shows the accuracy rates obtained by training the K-Nearest Neighbors (KNN) algorithm on the ‘Breast Cancer Wisconsin (Diagnostic)’ dataset with different number of features. The inferences observed from the data in Table 2 are as follows:

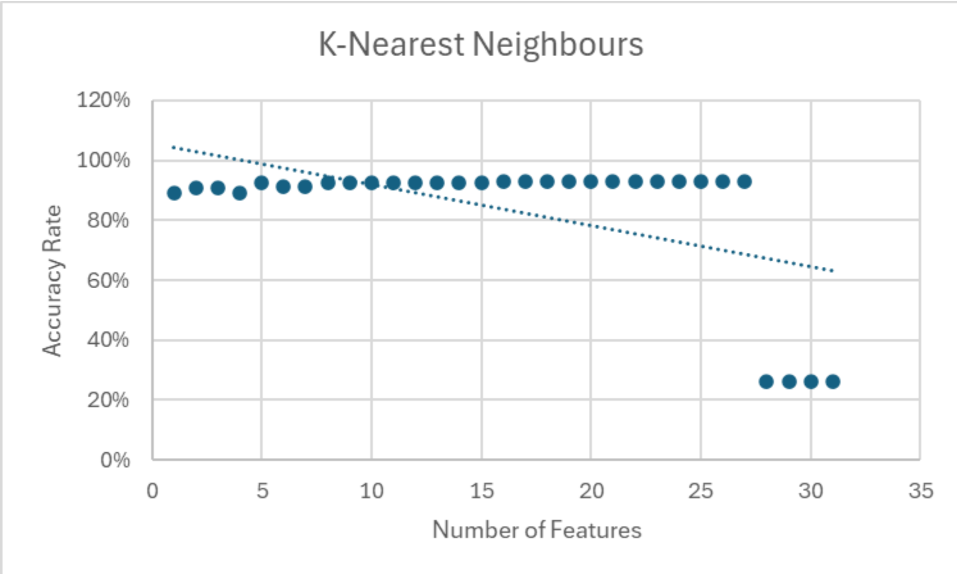
- **The number of features increases the accuracy rate up to a certain point:** A constant accuracy rate of almost 93% was obtained for the number of features between 16-27. These observed experimental results show that features in the range of 16-27 provide the most appropriate information for the K-Nearest Neighbors model.
- **Increasing the number of features does not always increase the accuracy rate:** When the number of features increases to 28, a serious decrease in the accuracy rate is observed. This may be an indicator of a phenomenon known as the curse of dimensionality. Too many features cause the model to become more complex and over-fit, reducing the ability to generalize.
- **A small number of features decreases the accuracy rate:** It is observed that the accuracy rate decreases when the number of features is below 8. This shows that the model does not have enough information to classify correctly at feature numbers below 8.

Table 2. Accuracy rates of K-Nearest Neighbors model

Number of Features	K-Nearest Neighbours	Number of Features	K-Nearest Neighbours	Number of Features	K-Nearest Neighbours
16	92,9762%	27	92,9762%	7	91,0432%
17	92,9762%	8	92,6284%	2	90,6924%
18	92,9762%	9	92,6284%	3	90,6924%
19	92,9762%	10	92,6284%	4	89,2794%
20	92,9762%	11	92,6284%	1	89,1071%
21	92,9762%	12	92,6284%	28	26,2500%
22	92,9762%	13	92,6284%	29	26,2500%
23	92,9762%	14	92,6284%	30	26,2500%
24	92,9762%	5	92,6253%	31	26,2500%
25	92,9762%	15	92,4499%		
26	92,9762%	6	91,0432%		

Source: Table created by the author

The accuracy rate trend is also shown on the graph of accuracy rates corresponding to increasing number of features (Graph 1), which was created using the decreasing accuracy rate and increasing number of features data given in**Hata! Başvuru kaynağı b ulunamadı.** Table 2.



Graph 1. Accuracy rates of K-Nearest Neighbors model
Source: Graph created by the author

4.3.3 Gaussian NB Model Accuracy Rates

Table 3 shows the accuracy rates obtained by training the Gaussian Naive Bayes (GNB) classification algorithm using different number of features. When Table 3 is analyzed:

- **Generally high accuracy rates were obtained:** Accuracy rates above 90% were achieved for most of the number of features. This shows that the GNB algorithm is a very successful classifier for this dataset.
- **Increasing the number of features does not always increase the accuracy rate:** The highest accuracy rates were achieved with the number of features being 3, 6, 11, 23. Adding more features decreased the accuracy rate in some cases. This is an indication of a situation called the curse of dimensionality. In such cases, too many features cause the model to become more complex and cause overfitting, reducing its ability to generalize.
- **It is seen that a small number of features can also decrease the accuracy rate:** When the number of features is

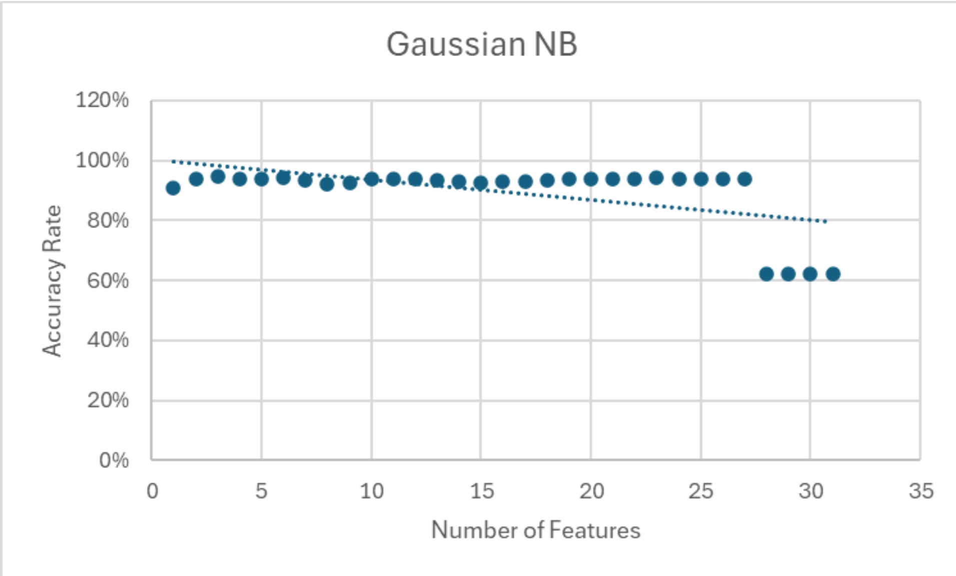
reduced to 1, the accuracy rate drops significantly. This indicates that the model does not have enough information for an accurate classification.

Table 3. Accuracy rates of Gaussian NB model

Number of Features	**Gaussian NB**	**Number of Features**	**Gaussian NB**	**Number of Features**	**Gaussian NB**
3	94,5551%	21	93,8534%	16	92,8008%
6	94,3797%	22	93,8534%	15	92,6253%
23	94,2043%	5	93,8534%	9	92,4499%
11	94,0320%	24	93,8534%	8	91,9236%
2	94,0288%	10	93,6779%	1	90,8678%
4	94,0288%	12	93,6748%	28	62,0395%
25	94,0288%	13	93,5025%	29	62,0395%
26	94,0288%	7	93,3271%	30	62,0395%
27	94,0288%	18	93,3271%	31	62,0395%
19	93,8534%	17	93,1516%		
20	93,8534%	14	92,8008%		

Source: Table created by the author

The accuracy rate trend is also shown on the graph of accuracy rates corresponding to increasing number of features (Graph 2), which was created using the decreasing accuracy rate and increasing number of features data given in Table 3.



Graph 2. Accuracy rates of Gaussian NB model
Source: Graph created by the author

4.3.4 Neural Network Model Accuracy Rates

Table 4 shows the accuracy rates obtained by training a neural network with different number of features. When Table 4 is analyzed, the following conclusions can be made:

- Accuracy Rate Variation: The increase in the number of features did not cause a continuous increase or decrease in the accuracy rate. Instead, a trend was observed where the accuracy rate peaked at a certain number of features and then decreased.
- Optimal Number of Features: In the Table 4, the highest accuracy rates were obtained at feature numbers 1, 8, 9, 11, 12 and 13. This suggests that the optimal number of features for this dataset may be in this range.
- Curse of Dimensionality: Too much increase in the number of features (27 and above) caused a serious decrease in accuracy. This is an indication of what is called the 'curse of dimensionality'. Too many features cause the model to

become more complex and lead to overfitting, which significantly reduces its generalization ability.

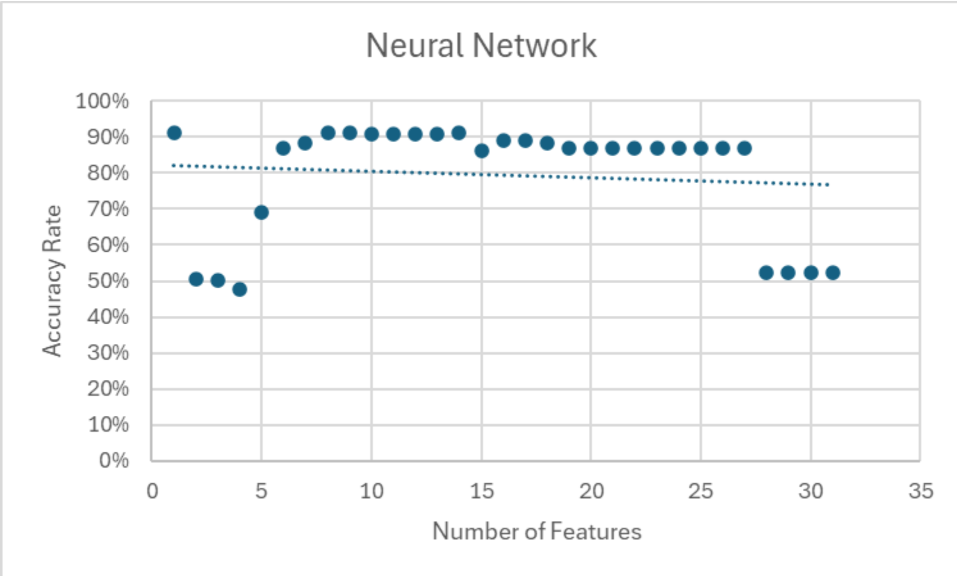
- The Effect of a Small Number of Features: If the number of features is very low (2, 3, 4, 5), the accuracy rate has decreased considerably. This serious decrease in the accuracy rate shows that the model does not have enough information to make the correct classification.

Table 4. Accuracy rates of Neural Network model

Number of Features	Neural Network	Number of Features	Neural Network	Number of Features	Neural Network
14	91,2030%	7	88,2237%	15	86,2625%
1	91,0401%	6	87,0019%	5	69,0758%
9	91,0276%	21	86,9799%	31	52,3465%
8	91,0276%	20	86,9799%	30	52,3465%
13	90,8521%	19	86,9799%	29	52,3465%
12	90,8521%	27	86,8045%	28	52,3465%
11	90,8521%	26	86,8045%	2	50,4292%
10	90,8521%	25	86,8045%	3	50,0658%
17	88,9317%	24	86,8045%	4	47,7851%
16	88,9317%	23	86,8045%		
18	88,4054%	22	86,8045%		

Source: Table created by the author.

The accuracy rate trend is also shown on the graph of accuracy rates corresponding to increasing number of features (Graph 3), which was created using the decreasing accuracy rate and increasing number of features data given in Table 4.



Graph 3. Accuracy Rates of Neural Network Model
Source: Graph created by the author

4.3.5 SVM-SVC Model Accuracy Rates

Table 5 shows the accuracy rates obtained by training the SVM-SVC (Support Vector Machine - Support Vector Classification) algorithm using different numbers of features. When Table 5**Hata! Başvuru kaynağı bulunamadı.** is analyzed, the f ollowing conclusions can be made:

- High and Stable Accuracy Rates: When the number of features used is less than 16, the accuracy rate remains quite high and stable at around 87%. This shows that the SVM-SVC algorithm is a powerful classifier for this dataset.
- Negative Effect of Increasing the Number of Features: In the case that the total number of features is 28, a serious decrease in the accuracy rate is observed. This is an indication of what is called the curse of dimensionality. Too many features cause the model to become more complex and overfit, reducing its generalization ability.
- Optimal Number of Features: The results obtained for determining the number of features in the dataset indicate

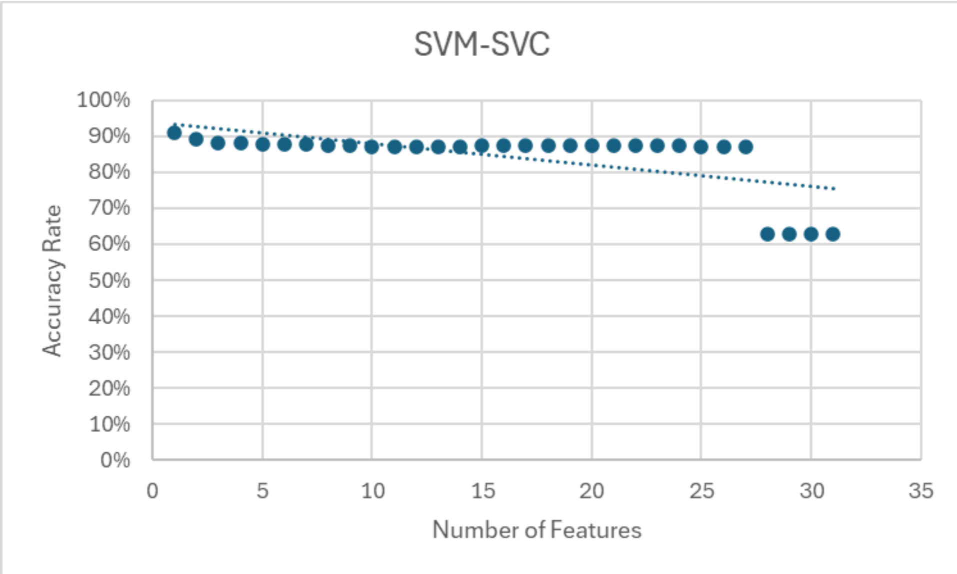
that the optimal value is approximately 15. However, this assertion may not always be true. This is bacause, optimal number of features may vary depending on the data patterns of the dataset used for evaluating the SVM-SVC model.

Table 5. Accuracy rates of SVM-SVC model

Number of Features	SVM-SVC	Number of Features	SVM-SVC	Number of Features	SVM-SVC
1	90,8647%	17	87,3496%	13	87,1742%
2	89,2826%	18	87,3496%	14	87,1742%
3	88,2299%	19	87,3496%	25	87,1742%
4	88,2299%	20	87,3496%	26	87,1742%
6	87,8759%	21	87,3496%	27	87,1742%
7	87,8759%	22	87,3496%	28	62,7412%
5	87,7036%	23	87,3496%	29	62,7412%
8	87,3496%	24	87,3496%	30	62,7412%
9	87,3496%	10	87,1742%	31	62,7412%
15	87,3496%	11	87,1742%		
16	87,3496%	12	87,1742%		

Source: Table created by the author.

The accuracy rate trend is also shown on the graph of accuracy rates corresponding to increasing number of features (Graph 4), which was created using the decreasing accuracy rate and increasing number of features data given in Table 5**Hata! Başvuru k aynağı bulunamadı.**.



Graph 4. Accuracy Rates of SVM-SVC Model
Source: Graph created by the author

4.3.6 SVM-NuSVC Model Accuracy Rates

Table 6**Hata! Başvuru kaynağı bulunamadı.** shows the a ccuracy rates obtained by training the SVM-NuSVC (Support Vector Machine - Nu-Support Vector Classification) algorithm using different numbers of features. When Table 6**Hata! Başvuru kaynağı b ulunamadı.** is analyzed, the following conclusions can be made:

- High and Stable Accuracy Rates: When the number of features used is less than 16, the accuracy rate remains quite high and stable at around 87%. This shows that the SVM-NuSVC algorithm is a powerful classifier for this dataset.
- Negative Effect of Increasing the Number of Features: In the case that the total number of features is 28, a serious decrease in the accuracy rate is observed. This is an indicator of a situation called the curse of dimensionality. Too many features cause the model to become more complex and overfit, reducing its generalization ability.
- Optimal Number of Features: The results obtained for determining the number of features in the dataset indicate

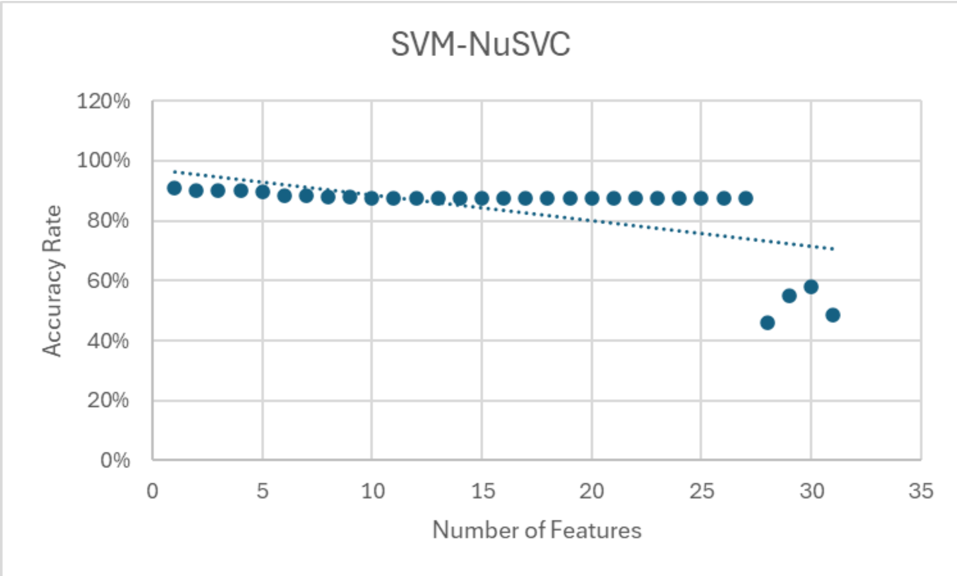
that the optimal value is approximately 15. However, this assertion may not always be true. This is bacause, optimal number of features may vary depending on the data patterns of the dataset used for evaluating the SVM-NuSVC model.

Table 6. Accuracy Rates of SVM-NuSVC Model

Number of Features	SVM-NuSVC	Number of Features	SVM-NuSVC	Number of Features	SVM-NuSVC
1	90,8647%	12	87,5251%	24	87,5251%
2	90,1598%	13	87,5251%	25	87,5251%
3	89,9843%	14	87,5251%	19	87,3496%
4	89,9843%	15	87,5251%	26	87,3496%
5	89,4580%	16	87,5251%	27	87,3496%
6	88,5777%	17	87,5251%	30	57,7851%
7	88,4023%	18	87,5251%	29	54,9781%
8	87,8759%	20	87,5251%	31	48,5307%
9	87,8759%	21	87,5251%	28	45,8553%
10	87,7005%	22	87,5251%		
11	87,5251%	23	87,5251%		

Source: Table created by the author.

The accuracy rate trend is also shown on the graph of accuracy rates corresponding to increasing number of features (Graph 5**Hata! Başvuru kaynağı bulunamadı.**), which was created u sing the decreasing accuracy rate and increasing number of features data given in Table 6**Hata! Başvuru kaynağı bulunamadı.**.



Graph 5. Accuracy rates of SVM-NuSVC model
Source: Graph created by the author

4.4 Evaluation of Hypotheses

This study analyses the accuracy rates of various machine learning models obtained using the 'Breast Cancer Wisconsin (Diagnostic)' dataset, which contains real-life data specified in Section 3.2. The results obtained by these models using different numbers of features are also examined in detail. Based on these results, the hypotheses presented in Section 3.1 are evaluated as follows:

H1: Feature selection improves the accuracy of machine learning models.

This hypothesis is supported when the accuracy rates obtained from training various models with different number of features are analyzed. The reason for the support of hypothesis H1 is that the results described below are obtained for the relevant models:

- **K-Nearest Neighbors (KNN) Model:** Exhibits a high degree of accuracy, with an approximate 93% accuracy rate observed for feature sets between 16 and 27. However, a notable decline in accuracy is evident when the number of features exceeds 28.

- **Gaussian NB Model:** Demonstrates the highest accuracy rates when utilizing 3, 6, 11 and 23 features. However, the addition of further features resulted in a decline in accuracy in certain instances.
- **Neural Network Model:** The highest accuracy rates were achieved with 1, 8, 9, 11, 12 and 13 features. Increasing the number of features caused a decrease in the accuracy rate.
- **SVM-SVC Model:** In cases with up to 15 features, the accuracy rate was found to be around 87%. When the number of features reached 28, a significant decrease in the accuracy rate was observed.
- **SVM-NuSVC Model:** When the number of features used is less than 16, the accuracy rate remains quite high and constant around 87%. In the case that the total number of features is 28, a significant decrease in the accuracy rate is observed.

The results are consistent with the hypothesis that the selection of appropriate features can enhance the accuracy of machine learning models.

H2: Different feature selection methods provide different performance improvements in different machine learning models.

The results obtained lend further support to this hypothesis:

- The **Gaussian NB** model and the **Neural Network** model demonstrated high accuracy rates with specific feature combinations, whereas the **K-Nearest Neighbors** model exhibited a consistent accuracy rate within a defined range of features.
- The **SVM-SVC** and **SVM-NuSVC** models also demonstrated disparate performance with varying numbers of features. Notably, high accuracy rates were observed with approximately 15 features, but this accuracy declined with the addition of more features.

It is evident that different feature selection methods yielded varying performance improvements across different models.

H3: Reducing the number of features prevents overfitting of the model and increases its generalization ability.

The results obtained lend further support to this hypothesis.

- **K-Nearest Neighbors** Model: As the number of features increases (28 and above), a notable decline in the accuracy rate is evident, which is regarded as an indicator of overfitting.
- **Gaussian NB** Model and **Neural Network** Model: In instances where the number of features increased to an excessive degree, a decline in accuracy rates was observed, indicating that the model was exhibiting signs of overfitting and a concomitant reduction in its generalization ability.
- The **SVM-SVC** and **SVM-NuSVC** models also demonstrated a notable decline in accuracy at feature numbers exceeding 28. This aligns with the observations made in the preceding models, indicating that SVM models are prone to overfitting with the data structure of the dataset at feature numbers exceeding 28. Consequently, their generalization capacity is diminished.

The hypothesis that reducing the number of features prevents overfitting and increases generalization ability is consistent with the results obtained.

5. CONCLUSION AND DISCUSSION

This study demonstrates that feature selection is a crucial technique for enhancing the accuracy of machine learning models and preventing model overfitting. The analyses conducted using diverse machine learning models (K-Nearest Neighbors, Gaussian NB, Neural Network, SVM-SVC, and SVM-NuSVC) reveal that accurate feature selection markedly enhances model performance.

The evaluation of the hypotheses is done as follows:

- **H1: Feature selection improves the accuracy of machine learning models.**

    - **Supported.** The results demonstrate that the models, particularly the K-Nearest Neighbors and Neural Network models, achieve high accuracy rates when the correct feature combinations are selected. Conversely, when the incorrect feature selection is employed, a significant decline in accuracy rates is observed.
- **H2: Different feature selection methods provide different performance improvements in different machine learning models.**
    - **Supported.** The Gaussian NB and Neural Network models demonstrated high accuracy rates when certain feature combinations were employed, whereas the K-Nearest Neighbors model exhibited a constant accuracy rate within a defined range of features. Additionally, the SVM-SVC and SVM-NuSVC models exhibited varying performance with varying numbers of features.
- **H3: Reducing the number of features prevents overfitting of the model and increases its generalization ability.**
    - **Supported.** In the K-Nearest Neighbors model, a notable decline in the accuracy rate was observed with an increase in the number of features, which was identified as an indicator of overfitting.

This study elucidates the influence of feature selection on the efficacy of machine learning models. In particular, it was demonstrated that the selection of optimal feature combinations markedly enhances model accuracy and prevents overfitting. When analyzing the performance of diverse models with varying numbers of features, it was observed that each model exhibits a distinct optimal feature combination.

The results of the study demonstrate that the process of selecting features for machine learning models should be conducted with careful consideration to enhance their performance. In particular, it is crucial to determine the optimal number of features to achieve the highest accuracy while preventing overfitting.

In consideration of the evaluations detailed in Section 4.4, all three hypotheses tested in this study are supported. These results demonstrate that feature selection improves the effectiveness of machine learning models and augments their generalization ability by preventing overfitting.

The following list presents potential avenues for future research on this subject:

- **Comparison of Different Feature Selection Methods:** Compare the effects of different feature selection methods (e.g., L1 regularization, PCA) on the performance and determine the most appropriate method.
- **Investigation of Ensemble Learning Approaches:** To investigate the performance of ensemble models created by combining different feature selection methods and machine learning models.
- **Relationship between Feature Selection and Model Complexity:**
  To study the effects of feature selection on model complexity and optimize the performance of more complex models.
- **Feature Selection in Time Series Data:** To investigate the effects of feature selection methods on time series data and to improve the performance of models working with these data.
- **Feature Selection in Large Datasets:** To examine the performance of feature selection methods in large datasets and to determine the most appropriate methods to improve the accuracy of models working with these data.

REFERENCES

Akkaya, M. (2021). Utility: Theories and Models. In *Applying Particle Swarm Optimization* (pp. 3-14). Springer. doi:10.1007/978-3-030-70281-6_1

Albrechts, L., & Balducci, A. (2013, 09 01). Practicing Strategic Planning: In Search of Critical Features to Explain the Strategic Character of Plans. *disP - The Planning Review, 49*(3), pp. 16-27. doi:10.1080/02513625.2013.859001

Ali, M. Z., Abdullah, A., Zaki, A. M., Rizk, F. H., M.Eid, M., & El-Kenway, E. M. (2024, 01 18). Advances and Challenges in Feature Selection Methods: A Comprehensive Review. *Journal of Artificial Intelligence and Metaheuristics, 7*(1), pp. 67-77. doi:10.54216/JAIM.070105

Anuragi, A., Sisodia, D. S., & Pachori, R. B. (2024, 02 26). Mitigating the curse of dimensionality using feature projection techniques on electroencephalography datasets - an empirical review. *Artificial Intelligence Review, 57*, p. 75. doi:10.1007/s10462-024-10711-8

Aytekin, H. T. (2021, 06 30). Makine Öğreniminin Araştırmacıların Veri Analizi Bağlamında Potansiyel Önemi. *Ufuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10*(19), pp. 85-106.

Bankins, S., & Formosa, P. (2023, 02 11). The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work. *Journal of Business Ethics, 185*, pp. 725–740. doi:10.1007/s10551-023-05339-7

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019, 06 24). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of*

Sciences (PNAS), 116(32), pp. 15849-15854. doi:10.1073/pnas.1903070116

Bellis, P., Magnusson, M., Nilsson, S., & Samuelsson, M. (2024, 06 27). Accelerating Innovation: Capabilities and Systems Beyond Organizational Boundaries. *Research-Technology Management, 67*(4), pp. 9-13. doi:10.1080/08956308.2024.2352688

Buyukkececi, M., & Okur, M. C. (2023). A Comprehensive Review of Feature Selection and Feature Selection Stability. *Gazi University Journal of Science, 36*(4), pp. 1506-1520. doi:10.35378/gujs.993763

CCPA. (2018). *California Consumer Privacy Act of 2018 (CCPA)*. Retrieved from California Legislative Information: https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5

Chen, O., Paas, F., & Sweller, J. (2023, 06 02). A Cognitive Load Theory Approach to Defining and Measuring Task Complexity Through Element Interactivity. *Educational Psychology Review, 35*(2), p. 63. doi:10.1007/s10648-023-09782-w

Crawford, J., & Jabbour, M. (2024). The relationship between enterprise risk management and managerial judgement in decision-making: A systematic literature review. *International Journal of Management, 26*(1), 110-136. doi:10.1111/ijmr.12337

Davis, F. D. (1989, 09). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly, 13*(3), pp. 319-340. doi:10.2307/249008

Demirer, M., Hernández, D. J., Li, D., & Peng, S. (2024, 02). Data, Privacy Laws and Firm Production: Evidence from the GDPR.

National Bureau of Economic Research, 32146. doi:10.3386/w32146

Dorini, G., Kapelan, Z., & Azapagic, A. (2011, 02). Managing uncertainty in multiple-criteria decision making related to sustainability assessment. *Clean Technologies and Environmental Policy, 13*(1), pp. 133-139. doi:10.1007/s10098-010-0291-7

Drèze, J., & Stern, N. (1987). Chapter 14 The theory of cost-benefit analysis. In A. J. Auerbach, & M. Feldstein (Eds.), *Handbook of Public Economics* (Vol. 2, pp. 909-989). Elsevier. doi:10.1016/S1573-4420(87)80009-5

Ducato, R. (2020, 06). Data protection, scientific research, and the role of information. *Computer Law & Security Review*, p. 105412. doi:10.1016/j.clsr.2020.105412

ECOA. (1974, 10 28). *The Equal Credit Opportunity Act*. Retrieved from U.S. Department of Justice - Civil Rights Department: https://www.justice.gov/crt/equal-credit-opportunity-act-3

Friedman, B., Hendry, D. G., & Borning, A. (2017, 01 01). A Survey of Value Sensitive Design Methods. *Foundations and Trends in Human-Computer Interaction, 11*(2), pp. 63-125. doi:10.1561/1100000015

Gao, L., & Guan, L. (2023, 10). Interpretability of Machine Learning: Recent Advances and Future Prospects. *IEEE MultiMedia, 30*(4), pp. 105-118. doi:10.1109/MMUL.2023.3272513

GDPR. (2018, 05 23). *General Data Protection Regulation (GDPR)*. Retrieved from General Data Protection Regulation (GDPR) - Legal Text: https://gdpr-info.eu/

Gonzalez, C. (2014). Decision-Making: A Cognitive Science Perspective. In S. E. Chipman (Ed.), *The Oxford Handbook of Cognitive Science.* Oxford University Press. doi:10.1093/oxfordhb/9780199842193.013.6

Gstrein, O. J., & Beaulieu, A. (2022, 1 29). How to protect privacy in a datafied society? A presentation of multiple legal and conceptual approaches. *Philosophy and technology, 35*(1), p. 3. doi:10.1007/s13347-022-00497-4

Guan, X., & Burton, H. (2022, 12). Bias-variance tradeoff in machine learning: Theoretical formulation and implications to structural engineering applications. *Structures, 46*, pp. 17-30. doi:10.1016/j.istruc.2022.10.004

Guerra-García, C., Nikiforova, A., Jiménez, S., Perez-Gonzalez, H. G., Ramírez-Torres, M., & Ontañon-García, L. (2023). ISO/IEC 25012-based methodology for managing data quality requirements in the development of information systems: Towards Data Quality by Design. *Data & Knowledge Engineering, 145*. doi:10.1016/j.datak.2023.102152

Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014, 08). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics, 154*, pp. 72-80. doi:10.1016/j.ijpe.2014.04.018

Howard, R. A. (1966, 08). Information Value Theory. *IEEE Transactions on Systems Science and Cybernetics, 2*(1), pp. 22-26. doi:10.1109/TSSC.1966.300074

Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., . . . Yi, X. (2020, 08). A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review, 37*, pp. 1-35. doi:10.1016/j.cosrev.2020.100270

ISO/IEC 23053:2022. (2022, 06). Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML). ISO.

ISO/IEC 23894:2023. (2023, 02). Information technology - Artificial intelligence Guidance on risk management. ISO.

ISO/IEC 42001:2023. (2023, 12). Information technology - Artificial intelligence Management system. ISO.

Jeffery, N. (2009). *Stakeholder Engagement: A Road Map to Meaningful Engagement.* Doughty Centre for Corporate Responsibility at the Cranfield School of Management. Retrieved from http://hdl.handle.net/1826/3801

Jiang, Z., Che, J., He, M., & Yuan, F. (2023, 02). A CGRU multi-step wind speed forecasting model based on multi-label specific XGBoost feature selection and secondary decomposition. *Renewable Energy, 203*, pp. 802-827. doi:10.1016/j.renene.2022.12.124

Kaggle. (1995, 10 31). Breast Cancer Wisconsin (Diagnostic) Data Set. Retrieved 07 25, 2024, from https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

Kalyuga, S. (2011, 01). Informing: A Cognitive Load Perspective. (T. G. Gill, Ed.) *Informing Science: the International*

Journal of an Emerging Transdiscipline, 14(1), pp. 33-45. doi:10.28945/1349

Kelly, M., Longjohn, R., & Nottingham, K. (2024, 07 25). Retrieved from The UCI Machine Learning Repository: https://archive.ics.uci.edu

Leopizzi, R. (2023). Stakeholder Engagement. In S. Idowu, R. Schmidpeter, N. Capaldi, L. Zu, M. Del Baldo, & R. Abreu (Eds.), *Encyclopedia of Sustainable Management* (pp. 3099–3103). Springer, Cham. doi:10.1007/978-3-031-25984-5_172

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017, 12 06). Feature Selection: A Data Perspective. *ACM Computing Surveys, 50*(6), pp. Aticle 94, 1-45. doi:10.1145/3136625

Macauley, M. K. (2005). The Value of Information: A Background Paper on Measuring the Contribution of Space-Derived Earth Science Data to National Resource Management. *Resources for the Future*, Discussion Papers 10839. doi:10.22004/ag.econ.10839

McCausland, T. (2024, 06 27). Accelerating Innovation. *Research-Technology Management, 67*(4), pp. 87-90. doi:10.1080/08956308.2024.2352997

Odek, R., & Oluoch, J. (2023, 02 10). Cost Benefit Analysis Origin And Applicability Of Its Recent Advances:. *Management and Economics Research Journal, 9*(1), p. 9900075. doi:10.18639/MERJ.2023.9900075

O'Sullivan, S., Nevejans, N., Allen, C., Blyth, A., Leonard, S., Pagallo, U., . . . Ashrafian, H. (2018, 10 26). Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery, 15*(1), p. e1968. doi:10.1002/rcs.1968

Parlak, B., & Uysal, A. K. (2021, 04 13). A novel filter feature selection method for text classification: Extensive Feature Selector. *Journal of Information Science, 1*, pp. 59-78. doi:10.1177/0165551521991037

Petropoulos, F., Laporte, G., Aktas, E., Alumur, S. A., Archetti, C., Ayhan, H., . . . Zhao, X. (2024). Operational Research: methods and applications. *Journal of the Operational Research Society, 75*(3), pp. 423-617. doi:10.1080/01605682.2023.2253852

Porath, U. (2023). Advancing Managerial Evolution and Resource Management in Contemporary Business Landscapes. *Modern Economy, 14*, pp. 1404-1420. doi:10.4236/me.2023.1410072

Prabhod, K. J. (2024, 04 16). The Role of Artificial Intelligence in Reducing Healthcare Costs and Improving Operational Efficiency. *Quarterly Journal of Emerging Technologies and Innovations, 9*(2), pp. 47–59. Retrieved from https://vectoral.org/index.php/QJETI/article/view/111

Steiner, G. A. (1979). *Strategic Planning.* New York: THE FREE PRESS.

Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993, 07 29). Nuclear feature extraction for breast tumor diagnosis. *Electronic imaging*, p. Corpus ID: 14922543. doi:10.1117/12.148698

Sweller, J. (2004). Instructional Design Consequences of an Analogy between Evolution by Natural Selection and Human Cognitive Architecture. *Instructional Science, 32*, pp. 9-31. doi:10.1023/B:TRUC.0000021808.72598.4d

Sweller, J. (2011). CHAPTER TWO - Cognitive Load Theory. In J. P. Mestre, & B. H. Ross (Eds.), *Psychology of Learning and Motivation* (Vol. 55, pp. 37-76). Academic Press. doi:10.1016/B978-0-12-387691-1.00002-8

Takemura, K. (2014). *Behavioral Decision Theory.* Springer Japan. doi:10.1007/978-4-431-54580-4

Tsirtsis, S., Tabibian, B., Khajehnejad, M., Singla, A., Schölkopf, B., & Gomez-Rodriguez, M. (2024, 02). Optimal Decision Making Under Strategic Behavior. *Management Science*. doi:10.1287/mnsc.2021.02567

Van Merrienboer, J. J., & Ayres, P. (2005, 09 01). Research on cognitive load theory and its design implication for E-learning. *Educational Technology Research and Development: ETR & D, 53*(3), pp. 5-13. doi:10.1007/BF02504793

Velliangiri, S., Alagumuthukrishnan, S., & Joseph, S. I. (2019). A Review of Dimensionality Reduction Techniques for Efficient Computation. *International Conference On Recent Trends In Advanced Computing 2019. 165*, pp. 104-111. Elsevier. doi:10.1016/j.procs.2020.01.079

Wang, P. (2019, 08). On Defining Artificial Intelligence. (D. Monett, C. W. Lewis, & K. R. Thorisson, Eds.) *Journal of Artificial General Intelligence, 10*(2), pp. 1-37. doi:10.2478/jagi-2019-0002

Wang, Y., & Ruhe, G. (2007). The Cognitive Process of Decision Making. *International Journal of Cognitive Informatics and Natural Intelligence, 1*(2), pp. 73-85. doi:10.4018/jcini.2007040105

Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1995, 10 31). Breast Cancer Wisconsin (Diagnostic). *UCI Machine Learning Repository*. doi:10.24432/C5DW2B

BÖLÜM VII

Early Detection of Parkinson's Disease through Machine Learning

Seda DEMİR[1]
Harun SELVİTOPİ[2]

Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder characterized by both motor and non-motor symptoms, affecting 1% of individuals aged 60 and older. This disease develops due to the loss of dopamine-producing cells in the brain, impacting motor functions and reducing the patient's quality of life. Studies on Parkinson's disease show that the disease typically starts with motor symptoms such as tremors, rigidity, and bradykinesia (slowness of movement). However, before the onset of motor symptoms, some non-motor symptoms may appear, including loss of smell, sleep disturbances, and constipation. These symptoms can emerge in the

[1] Research assistant., Erzurum Technical University, Department of Mathematics, Faculty of Science, Erzurum/Türkiye, Orcid: 0000-0003-0655-9326, seda.demir@erzurum.edu.tr.
[2] Associate professor, Erzurum Technical University, Department of Mathematics, Faculty of Science, Erzurum/Türkiye, Orcid: 0000-0001-5958-7625, harun.selvitopi@erzurum.edu.tr.

preclinical (early) stages of the disease and play a critical role in diagnosis (Jankovic, 2008:369; Braak et al., 2003: 124).

In recent years, the use of machine learning (ML) techniques for the early diagnosis and more accurate assessment of Parkinson's disease has become increasingly widespread. Machine learning has great potential in the healthcare field due to its ability to extract meaningful information from data and later use this information for prediction. ML applications in the diagnosis of Parkinson's disease are conducted on a wide range of data types, including written patterns, movement data, neuroimaging, voice, and biomarkers. The combination of these data types may enable more accurate diagnoses in the early stages of the disease (Cherubini et al., 2014a: 266; Segovia et al., 2019:1). For example, voice data is another commonly used biomarker in Parkinson's disease diagnosis, as speech impairments can be detected even before the onset of motor symptoms (Sakar et al., 2013:829).

ML methods have shown highly successful results in analyzing such biomarkers. Machine learning techniques not only analyze motor symptoms but also assess the cognitive and behavioral symptoms of Parkinson's disease (PD). However, non-motor symptoms are often not used independently in the diagnosis of the disease because these symptoms can vary from patient to patient and are difficult to evaluate (Zesiewicz et al., 2006:581). Nevertheless, some non-motor symptoms, particularly changes in voice, may support the diagnosis in the early stages of the disease (Postuma et al., 2015:1592). Therefore, machine learning techniques, supported by voice, written data, and other biomarkers, have great potential in the early diagnosis of PD through multi-modal data analysis methods. Machine learning is not limited to a single data type but also examines the combination of different data sources. For example, the integration of different neuroimaging techniques such as magnetic resonance imaging (MRI) and single-photon emission computed tomography (SPECT) data can enable more accurate diagnosis of the disease (Cherubini et al., 2014b:1216; Wang et al., 2017:222). The integration of data across these multiple modalities allows for a more comprehensive evaluation of patients and aids in earlier diagnosis. Analyses using machine learning play a crucial role in improving the clinical diagnosis of Parkinson's disease. Studies in this field have achieved high accuracy rates by utilizing data derived from various sources, including motor symptoms, voice impairments, handwriting patterns, and neuroimaging data. For instance, research analyzing motion and voice data has demonstrated that ML methods can enhance the accuracy of parameters derived from these data types (Yang et al., 2009:1; Wahid et al., 2015:1794; Anitha et al., 2020:251). Moreover, the application of machine learning techniques enables early diagnosis of PD, even in its preclinical stages, which can facilitate the initiation of treatment to slow the progression of the disease (Dorsey et al., 2018:939).

As a result, this study aims to provide a comprehensive review of machine learning methods used in the diagnosis of Parkinson's disease. The study will present a detailed summary of data sources, data types, employed ML models, and the results obtained, while also comparing the effectiveness of different machine learning techniques used in the early diagnosis of Parkinson's disease. Such a review can serve as a guide for future research and provide insights into how machine learning algorithms and new biomarkers can be utilized to make more accurate decisions in clinical applications.

1. Research Methodology

In this study, the UCI Parkinson's dataset was initially prepared. The data was then split into training and test sets. Subsequently, the data from both sets was processed and prepared for model input. The models used to classify Parkinson's disease, including XGBClassifier, Random Forest (RF), and Support Vector Machine (SVM), were trained using the training set. The accuracy of the models was evaluated based on their ability to classify the data in the test set. For the XGBClassifier model, k-fold cross validation was applied, and the results were compared. For the Random Forest model, accuracy values were calculated and compared both with and without SMOTE, as well as with feature selection algorithms. Finally, for the Support Vector Machine model, feature selection algorithms and k-fold cross validation were applied, and the accuracy values were compared.

1.1. Collection and Properties of Dataset

The dataset used in this study consists of features derived from the speech signals of 31 individuals, collected at the National Center for Voice and Speech in Denver, Colorado. Created by Max Little from the University of Oxford, the dataset was contributed to the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/174/parkinsons). It includes a series of biomedical voice measurements from 31 individuals, 23 of whom have Parkinson's disease (PD). Each column represents a specific voice measurement, while each row corresponds to one of the 195 voice recordings obtained from these individuals ("name" column). The primary purpose of the data is to distinguish individuals with PD from healthy individuals based on the "status" column, where 0 indicates healthy and 1 indicates PD. The class distribution of the dataset is illustrated in Figure 1, showing 48 healthy phonations and 147 PD phonations across the 31 individuals.

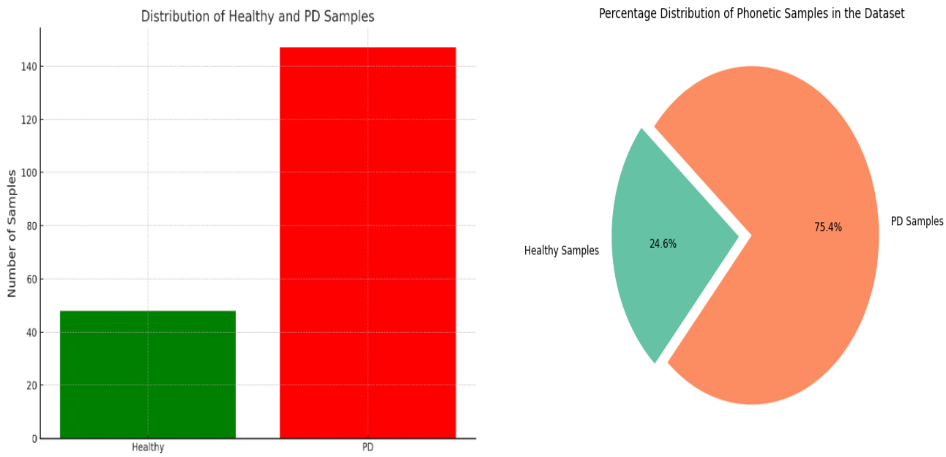


Figure 1: Distribution of phonetic samples by class in the dataset.

1.2. Data Pre-processing

Data pre-processing is a fundamental step for the successful application of machine learning models. This stage involves applying various techniques to transform raw data into a format suitable for the model. Raw data often contains missing values, imbalanced classes, features with different scales, or irrelevant variables. Therefore, the data pre-processing process aims to improve the quality of the data, enhancing the model's accuracy and generalization performance. The basic steps in data preprocessing include handling missing values, data scaling, elimination of class imbalances, feature selection, and data transformations. Missing values can be filled using statistical methods or advanced imputation algorithms. Features with varying scales can adversely affect the performance of certain machine learning algorithms; this issue is resolved using scaling techniques. Additionally, class imbalance problems are addressed using oversampling or undersampling methods to ensure fair learning across classes.

In this study, various data pre-processing techniques were applied to effectively train different models. For the Random Forest (RF) model, the Synthetic Minority Oversampling Technique (SMOTE) was used to balance the class distribution, and important features were selected using the Recursive Feature Elimination (RFE) method. For the XGBClassifier model, k-cross validation was applied to evaluate the model's generalization performance. For the Support Vector Machine (SVM) model, data were scaled using the Standard Scaler method, and the model's performance was tested using k-cross validation. Each of these processes generally improved the models' performance and enhanced the accuracy of the analysis results. However, in some cases, the application of certain data pre-processing techniques led to a decline in accuracy. This highlights the critical role of selecting appropriate data pre-processing steps and ensuring that the methods align with the structure of the dataset in determining the success of machine learning models.

1.2.1. SMOTE (Synthetic Minority Oversampling Technique)

SMOTE is a widely used data preprocessing technique to address class imbalance issues. It balances the dataset by oversampling the minority class, enabling the model to learn the minority class patterns more effectively. Synthetic samples for the minority class are generated through linear interpolation between existing minority examples in the dataset. This approach reduces the risk of overfitting to the majority class, which is a common problem in imbalanced datasets, and ensures fairer classification across classes. SMOTE plays a crucial role in improving performance metrics such as accuracy, F1 score, and AUC, especially in scenarios where class imbalance heavily skews model results.

1.2.2. Standard Scaler

Differences in feature scales can negatively impact the performance of machine learning models. Standard Scaler addresses this issue by standardizing features to have a mean of 0 and a standard deviation of 1. This method is particularly beneficial for distance-sensitive algorithms, such as Support Vector Machines and K-Nearest Neighbors, as it ensures consistent performance across features with varying scales. By normalizing feature values to a common scale, Standard Scaler enables models to treat all features equally, improving the overall accuracy and reliability of the predictions. Additionally, this scaling technique enhances model convergence and reduces training time.

1.2.3. RFE (Recursive Feature Elimination)

Not all features in a dataset contribute equally to the performance of a machine learning model. RFE is a feature selection technique that identifies the most important features by iteratively removing those with the least impact on model performance. Initially, the model is trained using all features, and the significance of each feature is evaluated. The least important feature is then eliminated, and the process repeats until the optimal subset of features is obtained. RFE not only enhances model performance but also improves interpretability and reduces computational complexity. This technique is particularly effective in high-dimensional datasets, enabling faster model training and mitigating overfitting risks.

1.2.4. k-Cross Validation

k-cross validation is an effective technique for assessing the generalization performance of a model. In this method, the dataset is divided into k equal parts, and the model is trained and tested k times. In each iteration, one part is used as the test set, while the remaining k-1 parts serve as the training set. This process ensures that every subset of the data is used for both training and testing across k iterations. k-cross validation provides insights into how the model performs on different data subsets and helps prevent overfitting. Additionally, it delivers more reliable estimates of generalization errors, making it invaluable for selecting the best-performing model. This method is particularly useful in cases where the dataset size is limited.

2. Classifier Models

2.1. XGBoost Classifier (XGBClassifier)

XGBoost, short for Extreme Gradient Boosting, is a robust and scalable machine learning algorithm commonly used for classification tasks. It belongs to the ensemble learning family, utilizing gradient boosting techniques where a series of decision trees are sequentially trained to minimize a loss function. XGBoost incorporates advanced features like regularization, parallel processing, and tree pruning, which enhance training efficiency and model performance. Its adaptability and speed make it suitable for diverse applications, including structured data, text, and image classification. The final prediction results from combining the outputs of all the individual trees in the ensemble.

2.2. Random Forest (RF)

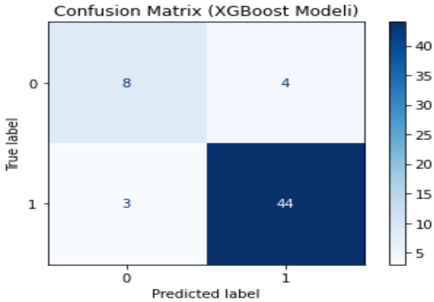
Random Forest is a powerful ensemble learning algorithm that trains multiple decision trees independently, with the final classification or regression result determined by averaging or taking the majority vote of the individual tree predictions. By selecting random subsets of data and features for each tree, it introduces diversity, which helps reduce overfitting and improves the model's generalization ability. Additionally, by using only a portion of features during the training of each tree, the correlation between features is reduced, leading to more reliable and stable predictions. These characteristics make Random Forest an effective tool for classification and regression, especially for large and complex datasets, offering high accuracy and robustness.

2.3. Support Vector Machine (SVM)

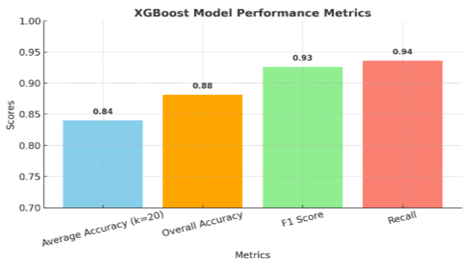
Support Vector Machine (SVM) is a powerful supervised learning algorithm primarily used for classification tasks, though it can also be applied to regression problems. The core idea behind SVM is to find the optimal hyperplane that best separates data points of different classes in a high-dimensional space. It aims to maximize the margin, which is the distance between the closest data points (support vectors) and the hyperplane. SVM can handle both linear and non-linear classification by using kernel functions, such as the radial basis function (RBF) kernel, to map data into a higher-dimensional space where a linear separator can be found. The algorithm is known for its robustness, especially in high-dimensional spaces, and its ability to perform well even with a small number of training samples. Despite its effectiveness, SVM can be computationally intensive, particularly with large datasets.

3. Results and Discussions

In this study, machine learning methods were applied to the UCI Parkinson's dataset to detect cardiovascular diseases, with performance measures including accuracy, precision, recall, and F1 score derived from the confusion matrix. Additionally, ROC curves and AUC values were calculated and compared.



Metric	Value
ROC AUC Score	0.8014
F1 Score	0.9263
Precision	0.9167
Recall	0.9362
Average Accuracy (k=5)	0.7590
Average Accuracy (k=10)	0.7771
Average Accuracy (k=15)	0.8256
Average Accuracy (k=20)	0.8400
Best k Value	20
Accuracy Score (Without Cross-Validation)	0.8814



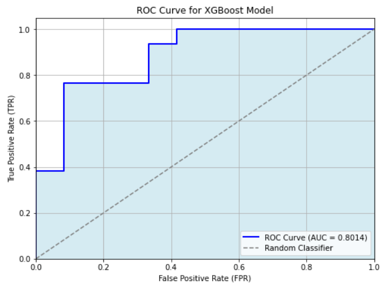


Figure 2: Performance metrics for the XGBClassifier model.

The bar chart above illustrates the performance metrics of the XGBoost model. The graph compares the highest average accuracy ($\%84.00, k = 20$) from cross-validation with overall accuracy ($\%88.14$), F1 score ($\%92.63$), and recall rate ($\%93.62$). It is observed that the model has a particularly strong and reliable predictive ability. When cross-validation was applied to the XGBoost model, the accuracy values showed a slight decrease compared to the non-cross-validation scenario. During the cross-validation process, the dataset was divided into different subsets to test the model's generalization ability. This method provides a more realistic evaluation of the model's performance, but the overall accuracy rate ($\%88.14$) was higher than the cross-validation accuracy ($\%84.00$) for $k = 20$. This suggests that the model might be affected by the different distributions of the data subsets during cross-validation.

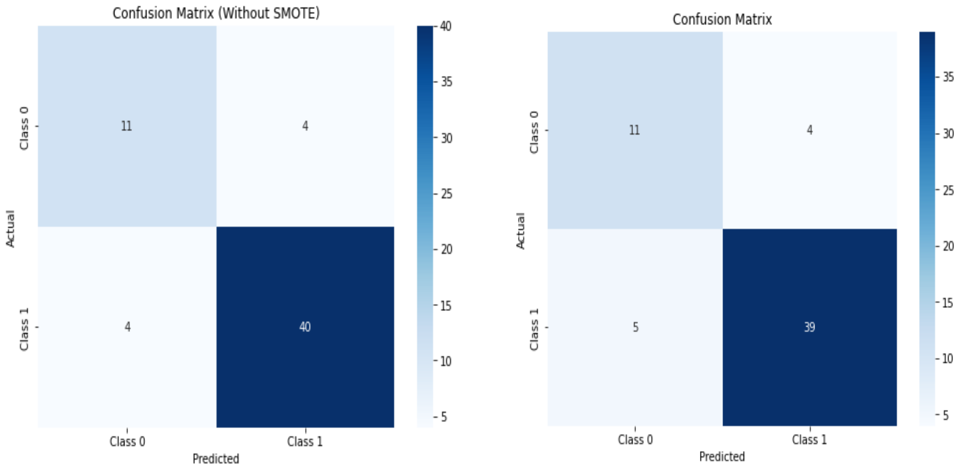


Figure 3: Confusion matrix with and without Smote applied for Random Forest model.

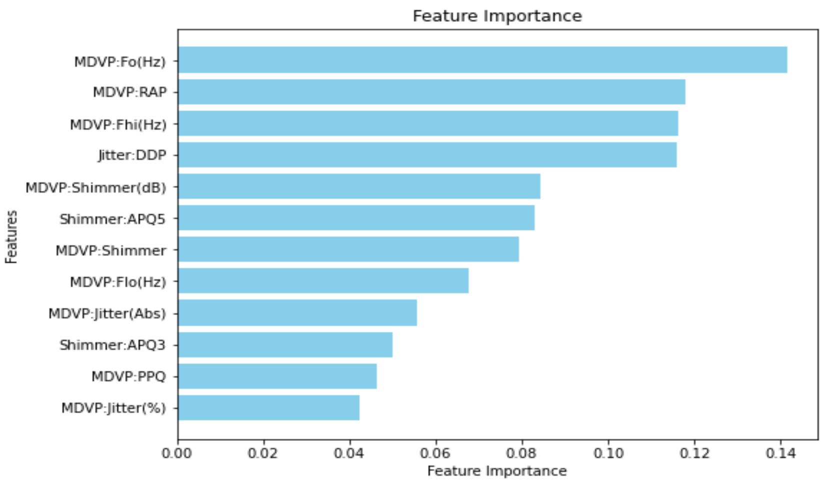


Figure 4: Feature levels importance for Random Forest model.

Table 1: Performance metrics comparasion for Random Forest model (Without SMOTE, With SMOTE and With Feature Selection)

Metrics	Without SMOTE	With SMOTE	With Feature Selection
Accuracy	0.86	0.85	0.86
Precision	0.91	0.91	0.89
Recall	0.91	0.89	0.93
F1-Score	0.91	0.90	0.91
AUC	0.9030	0.8902	0.80

Before applying SMOTE (Synthetic Minority Over-sampling Technique), the model achieved an accuracy of 86%. However, after applying SMOTE, the accuracy decreased slightly to 85%. Precision remained unchanged at 91% both before and after SMOTE. Recall dropped from 91% before SMOTE to 89% after SMOTE, but increased to 93% after feature selection. The F1-score decreased from 91% after feature selection to 90% with SMOTE. The AUC (Area Under the Curve) value showed a small decline, from 0.9030 before SMOTE to 0.8902 after SMOTE. After feature selection, the AUC further dropped to 0.80. These findings suggest that while SMOTE negatively impacted some performance metrics, feature selection helped partially recover the losses in performance.

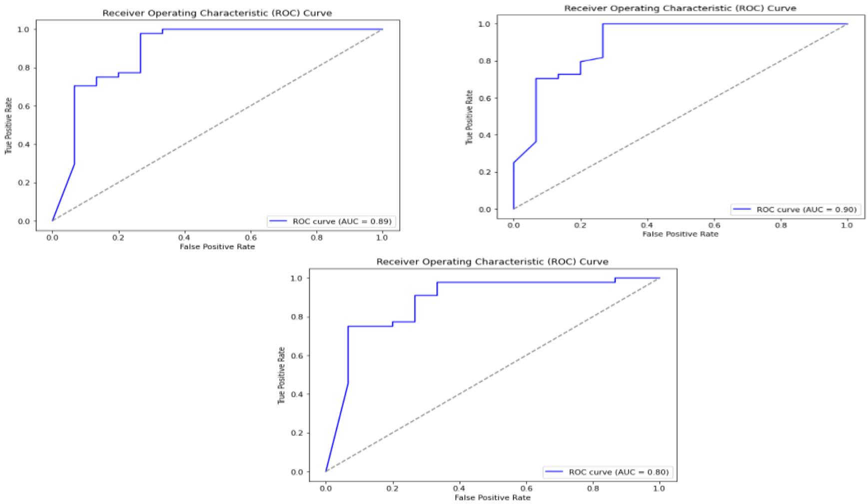


Figure 5: ROC curves and AUC values for Random Forest. The first column is the AUC value without applying SMOTE on the left, and the AUC value after applying Smote on the right. AUC value after applying SMOTE and feature selection in the second column.

Table 2: Performance metrics comparasion for Support vector machine model (Without RFE, With RFE and With StandardScaler).

Metric	SVM (Without RFE)
Accuracy	0.8205
Recall	0.9062
Precision	0.8788
F1 Score	0.8923
AUC	0.7411

Metric	SVM (With RFE)
Accuracy	0.7692
Recall	0.8438
Precision	0.8710
F1 Score	0.8571
AUC	0.7411

Metric	SVM (With StandardScaler)	Optimal k Value
Accuracy	0.8269	2
Recall	0.9580	17
Precision	0.8418	4
F1 Score	0.8928	9
AUC	0.8492	17

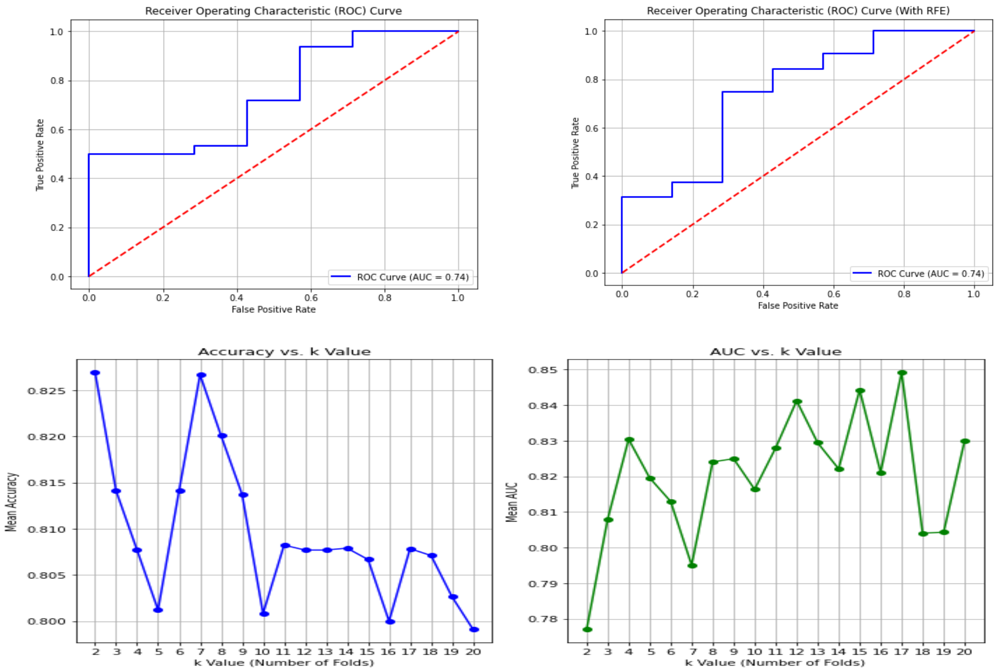


Figure 6: ROC curve and AUC value comparasion for Support vector machine model (Without RFE, With RFE and With StandardScaler).

The SVM model without RFE outperforms the model with RFE in terms of accuracy, recall, precision, and F1 score. This suggests that RFE may not improve the model's performance, and could even cause a decrease in these metrics. Specifically, the AUC values remain unchanged in both cases, indicating that RFE does not significantly affect the model's ability to distinguish between classes. On the other hand, the hyperparameter tuning through k-fold cross validation proves to be quite effective in optimizing the model based on different metrics. In particular, the highest average values for Recall and F1 Score were achieved with higher k values ($k = 17$ for Recall, $k = 9$ for F1 Score). These results suggest that accuracy and F1 score can be further improved with higher k values. However, Accuracy and AUC showed slight improvements at $k = 2$ and $k = 17$. In conclusion, RFE had a negative effect on the model's performance, reducing the accuracy and other metrics. However, hyperparameter optimization, particularly for Recall and F1 Score, significantly enhanced the model's performance and resulted in more accurate predictions. This demonstrates the critical role of selecting the right hyperparameters to improve the model's generalization ability.

4. Conclusion

In this study, 3 machine learning (ML) methods were applied to the UCI Parkinson's dataset. Before applying ML methods for data analysis, SMOTE, feature selection, and k-cross validation techniques were used. The performance of these methods was evaluated using various measurement methods and compared. The applied methods were compared both among themselves and with the literature. The results of the proposed methods in this study were observed to be better than those obtained in the literature.

5. References

Anitha, R., Nandhini, T. S. R. S., Raj, S. S., & Nikitha, V. (2020). Early detection of Parkinson's disease using machine learning. *IEEE Access*, 8, 147635-147646.

Braak, H., Ghebremedhin, E., Rüb, U., Bratzke, H., & Del Tredici, K. (2004). Stages in the development of Parkinson's disease-related pathology. *Cell and tissue research*, 318, 121-134.

Cherubini, A., Morelli, M., Nisticó, R., Salsone, M., Arabia, G., Vasta, R., Augimeri, A., Caligiuri, M. E., & Quattrone, A. (2014). Magnetic resonance support vector machine discriminates between Parkinson disease and progressive supranuclear palsy. *Movement Disorders*, 29(*2*), 266-269.

Cherubini, A., Nisticó, R., Novellino, F., Salsone, M., Nigro, S., Donzuso, G., & Quattrone, A. (2014). Magnetic resonance support vector machine discriminates essential tremor with rest tremor from tremor-dominant Parkinson disease. *Movement Disorders*, 29(*9*), 1216-1219.

Dorsey, E. R., Elbaz, A., Nichols, E., Abd-Allah, F., Abdelalim, A., Adsuar, J. C., et al. (2018). Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol*. 17, 939–953. doi: 10.1016/S1474-4422(18)30295-3.

Jankovic, J. (2008). Parkinson's disease: clinical features and diagnosis. *Journal of neurology, neurosurgery & psychiatry,* 79(*4*), 368-376.

Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., Obeso, J., Marek, K., Litvan, I., Lang, A. E., Halliday, G., Goetz, C. G., Gasser, T., Dubois, B., Chan, P., Bloem, B. R., Adler, C. H., & Deuschl, G. (2015). MDS clinical diagnostic criteria for Parkinson's disease. *Movement Disorders*, 30(*12*), 1591–1601.

Sakar, B. E., Isenkul, M. E., Sakar, C. O., Sertbas, A., Gurgen, F., Delil, S., Gurgen, F., & Kursun, O. (2013). Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE journal of biomedical and health informatics*, 17(*4*), 828-834.

Segovia, F., Górriz, J. M., Ramírez, J., Martínez-Murcia, F. J., & Castillo-Barnes, D. (2019). Assisted diagnosis of Parkinsonism based on the striatal morphology. *International Journal of neural systems,* 29(*9*), 1950011.

Yang, M., Zheng, H., Wang, H., & McClean, S. (2009, April). Feature selection and construction for the discrimination of neurodegenerative diseases based on gait analysis. *In 2009 3rd International Conference on Pervasive Computing Technologies for Healthcare* (pp. 1-7). IEEE.

Zesiewicz, T. A., Sullivan, K. L., & Hauser, R. A. (2006). Nonmotor symptoms of Parkinson's disease. *Expert review of neurotherapeutics*, 6(*12*), 1811-1822.

Wahid, F., Begg, R. K., Hass, C. J., Halgamuge, S., & Ackland, D. C. (2015). Classification of Parkinson's disease gait using spatial-temporal gait features. *IEEE journal of biomedical and health informatics*, 19(*6*), 1794-1802.

Wang, Z., Zhu, X., Adeli, E., Zhu, Y., Nie, F., Munsell, B., & Wu, G. (2017). Multi-modal classification of neurodegenerative disease by progressive graph-based transductive learning. *Medical image analysis*, 39, 218-230.

BÖLÜM VIII

Akıllı Şehirler: Bibliyometrik Analiz

Nevin AYDIN[1]

Giriş

Akıllı şehirler, vatandaşların yaşam kalitesini ve kentsel hizmetlerin verimliliğini artırmak için teknoloji ve verilerden yararlanır. Nesnelerin interneti (IoT) ve makine öğrenimi, büyük ölçekli veri toplama, analiz ve karar almaya yönelik akıllı şehirlerde çözüm üretme teknolojileri olarak ortaya çıkmıştır. Bu teknolojiler daha yaşanabilir, sürdürülebilir ve verimli şehirlerin yaratılmasında büyük potansiyele sahiptir. Buna karşılık, akıllı şehirlerin potansiyelini gerçekleştirmek için ele alınması gereken veri gizliliği, güvenlik ve etik konular ile ilgili zorluklar devam etmektedir.

Akıllı şehirler, Nesnelerin İnterneti (IoT) sensörleri, bağlı cihazlar, yapay zeka (AI) ve gerçek zamanlı olarak çeşitli kaynaklardan veri toplamak ve analiz etmek için büyük veri analitiği gibi çeşitli teknolojilerden faydalanır. Akıllı bir şehrin amacı, bu verileri ulaşım, enerji, su yönetimi, atık yönetimi ve kamu güvenliği

1 [1]Prof. Dr., Artvin Çoruh Üniversitesi, İİBF, İşletme Bölümü, nevin.aydin@gmail.com

gibi kentsel hizmetlerin verimliliğini ve etkinliğini artırmak için kullanmaktır. Şehirlerin geleceğinin teknolojik, sosyal ve çevresel eğilimlerin bir birleşimi olarak şekillendirilmesi gereklidir.

Akıllı şehir kavramları, kentleşmenin getirdiği zorluklara ve sürdürülebilir kentsel gelişime duyulan ihtiyaçtan ortaya çıkmıştır. Nesnelerin İnterneti, yapay zeka ve büyük veri analitiği gibi çeşitli teknolojilerin birlikte kullanımı, kentsel hizmetlerin verimliliğini artırmak, kaynak tüketimini azaltmak ve yaşam standardını yükseltmek gibi amaçları hedeflemektedir; bu vatandaşlar için daha iyi bir yaşam kalitesi yaratır. Teknolojilerin gelişimiyle, akıllı şehirler, akıllı tarım, akıllı endüstri, akıllı çiftçilik, akıllı sağlık hizmeti, akıllı trafik ve akıllı yayalar gibi çeşitli uygulamalardan veri toplayıp, yaşam standardını iyileştirmek için verileri analiz ederek entegre edebilirler (Wang vd., 2018).

Bibliyometrik Analiz

Bibliyometrik analiz, belirli bir konu üzerinde üretilen akademik yayınlar arasındaki ilişkilerin sayısal analizi olarak tanımlanır. Yayın veri tabanlarından elde edilen bibliyografik bilgiler aracılığıyla sayısal verileri kullanan bibliyometrik analiz, yazarlar, kurumlar ve ülkelerle ilgili atıf analizi, atıf grafikleri ve anahtar kelime grafikleri gibi sonuçları analiz etmek ve sunmak için bir yöntem olarak ortaya çıkar. Bibliyometrik analiz, belirli bir konu üzerinde akademik yayın üreten araştırmacıların etkisini ve araştırmacılar arasındaki etkileşimin boyutlarını ortaya çıkarmayı amaçlar. Bibliyometrik analiz, belirli bir alan veya disiplin içindeki kalıpları ve eğilimleri daha iyi anlamak için bibliyografik verilerin sistematik olarak incelenmesini içeren nicel bir araştırma yöntemidir. Bu yöntem, araştırmacıların araştırma alanındaki önemli yazarları, etkili yayınları ve ortaya çıkan konuları belirlemesini sağlar. Bibliyometrik analiz, bilimsel veri tabanlarından, dergilerden, konferans bildirilerinden ve diğer ilgili yayınlardan bibliyografik verilerin toplanmasını ve analiz edilmesini içerir. Genel olarak, bibliyometrik analiz, atıf sayıları, ortak yazarlık ağları, yayın kalıpları ve anahtar kelime analizi gibi temel metriklerin yazılım araçları veya platformları aracılığıyla belirlenmesine olanak tanır. Bibliyometrik analiz çeşitli araçlar ve teknikler aracılığıyla kolaylaştırılabilir. Bu araçlar, ilgili bilimsel yayınların toplanmasına, düzenlenmesine ve analiz edilmesine yardımcı olur. Bu araçlardan bazıları şunlardır: Scopus, Web of Science, Google Scholar.

Metodoloji

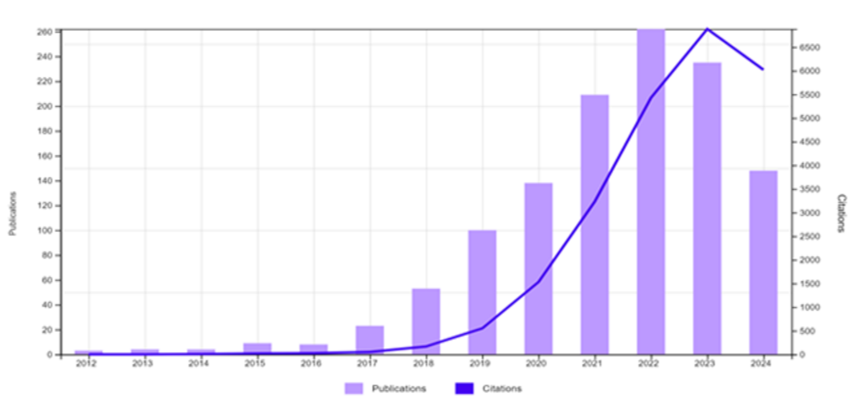
Bu çalışmada, 2012-2024 yılları arasında Web of Science veri tabanında "Akıllı şehirler", "Yapay Zeka" ve "Nesnelerin Interneti" anahtar sözcükleri kullanılarak yapılan bibliyometrik analiz sonucunda toplam 1375 sonuç elde edilmiştir. Çalışma 24 Ekim 2024 tarihinde yapılmıştır. Konuyla ilgili 2012 yılı ve 2024 yılları arasında yayınlanan çalışmalar ele alınmıştır. 1375 sonuç arasında literatürde "Akıllı şehirler", "Nesnelerin Interneti "Yapay Zeka" kavramlarıyla ilgili 915 makale, 259 işlemlem gören makale, 14 kitap bölümü, 12 editoryal materyal, 154 incelemede olan makale, 4 çekilen yayın, 21 erken erişim yayını, ve 1 yeni gelen özet olduğu tespit edilmiştir. VOSviewer programını kullanarak akıllı şehir kavramı üzerine akademik araştırmalar araştırılmış ve haritalanmıştır. Web of Sciences veri tabanından alınan veriler, yazar, atıf, dergi, ülke, organizasyon ve anahtar sözcükler temelinde analiz edilmiştir.

Nicel tekniklerin kullanılması, söz konusu araştırma alanındaki mevcut durumun ve gelişim eğilimlerinin tanımlanmasına olanak tanır. Sonuçlar, belirli bir dönemdeki ana araştırma yönleri, eğilimler ve yayın sayısındaki değişiklikler hakkında fikir verir.

Araştırma Sonuçları

Yıllara Göre Dağılım

Yayınların 2012-2024 veri aralığına dahil edildiği Şekil 1’de, sürekli artışlar ve azalışlarla birlikte süreksiz bir evrim gözlemlenmektedir. En üretken yıl 2022 ile en düşük yayın sayısına sahip yıl 2012 olduğu görülmektedir.



Şekil 1. 2012-2024 yılları arasındaki yayın ve atıfları göstermektedir.

Tablo 1. Yıllara Göre Yayın Sayısını Göstermektedir.

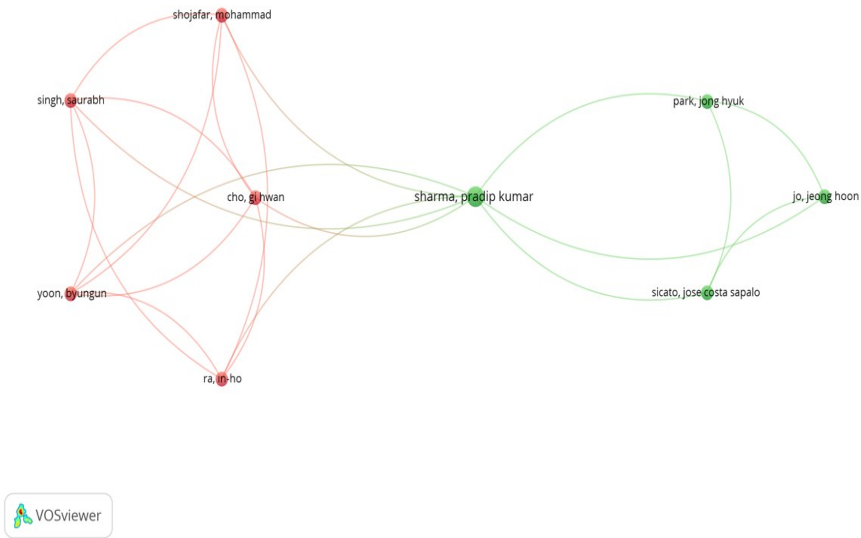
2024	**2023**	**2022**	**2021**	**2020**	**2019**	**2018**	**2017**	**2016**	**2015**	**2014**	**2013**	**2012**
195	276	300	231	157	106	54	25	11	9	4	4	3

Kaynak: Yazar tarafından düzenlenmiştir.

Ortak Yazar Analizi (Co-Authorship with Authors)

Ortak yazar analizi, yazarlar arasındaki iş birliği ağlarını tespit etmek için kullanılmaktadır (Kurnaz, 2021; Hırlak, 2024).

Bu analizde en az bir yayını ve en az bir atıfı bulunan yazarların listelenmesi baz alınmıştır. Bu kriter seçimi ile 187 yazardan 143'i incelemeye dahil edilmiştir. Daha büyük daireler ve harita etiketleri daha büyük önemi temsil etmektedir. Benzer renklere sahip olanlar aynı kümeye aittir (van Eck ve Waltman, 2010). Şekil 2'de yazarların konu hakkındaki yayın sayısına göre yoğunluğu ve yazarlar arası iş birliği ağı görülmektedir.



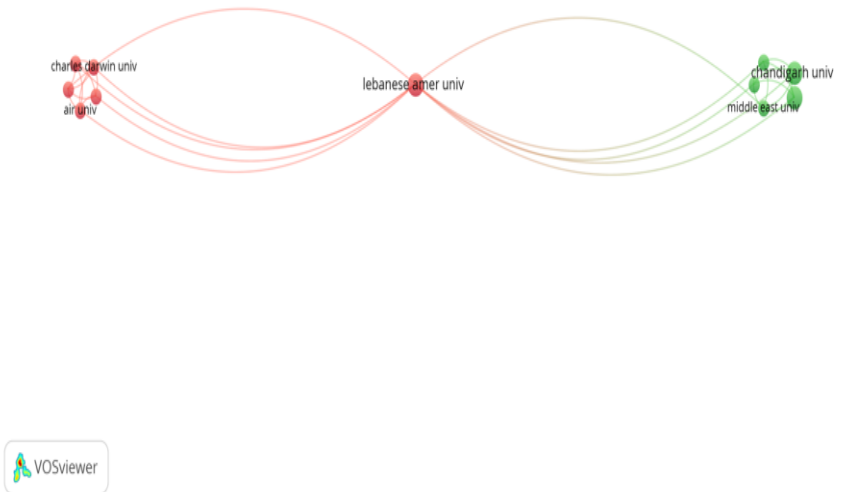
Şekil 2. Ortak Yazar Analizi Ağ Haritası

Yazarların ilişki durumunun iki alt kümede toplandığı görülmektedir. Her bir kümedeki yazarların birbiri ile çalışmaları bulunmaktadır. Bununla beraber bu analizle bu konu hakkında en çok yayın yapan ve en çok atıf alan yazarlar görülmektedir. Sayısal değer olarak bakılacak olduğunda konu hakkında en çok atıf alan ilk 10 yazar Tablo 2'de gösterilmektedir.

Tablo 2. Ortak Yazarlı Ağ Kümesi

Küme 1 (Kırmızı)	**Küme 2 (Yeşil)**
Cho, gi hwan	Jo, jeong hoon
Ra, ın-ho	Park, jong hyuk
Shojafar, mohammad	Sharma, pradip kumar
Sing, saurabh	Sicato, jose costa sapalo
Yoon, byungun	

Ortak Yazarlı Organizasyon Analizi (Co - authorship with Organizations)



Şekil 3. Ortak Yazarlı Organizasyon Analizi Haritası.

Organizasyonda en az 1 makale ve 1 atıf olsun, 133 organizasyondan 102 si bu eşiği gerçekleştiriyor. İki kümeden oluşuyor. Kırmızı ve yeşil kümeler Tablo 3'de gösterilmiştir. Tablo 4'de organizasyonların makale ve atıf sayıları görülmektedir. lebanese amer univ, 2 yayın, 41 atıf ile birinci; birla ınst technol & sci pilani 2 yayın, 32 atıf ile ikinci; univ amer 2 yayın, 48 atıf ile üçüncü sırada görülmektedir.

Tablo 3. Ortak Yazarlı Organizasyon Ağ Kümesi

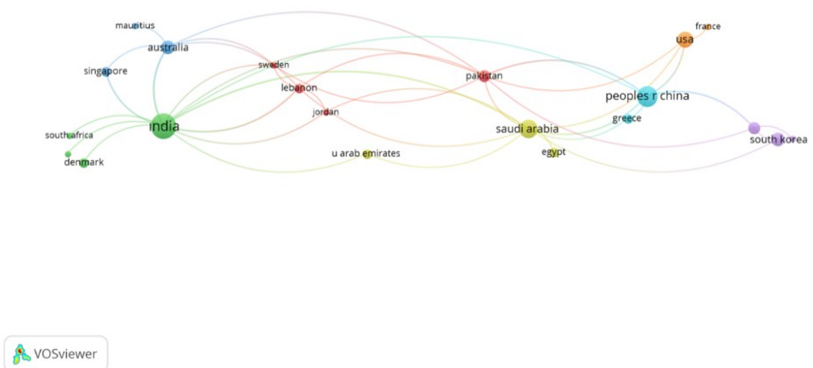
Küme 1 (Kırmızı)	Küme 2 (yeşil)
air univ	chandigarh univ
charles Darwin univ	chitkara univ
lebanese amer univ	lovely profess uni
linnaeus univ	middle east univ
symbiosis ınt deemed univ	princess nourah bint abdul
vellore ıns technol	

Tablo 4. Ortak Yazarlı Organizasyonda Makale ve Atıf sayısı Ağ Kümesi

Organizasyon	Makale	Atıf	Etki gücü
lebanese amer univ	2	41	10
birla ınst technol&sci pilani	2	32	9
univ amer	2	48	7
beijing univ technol	1	19	6
ccıs prince sultan univ	1	19	6
chandigarh univ	2	1	6
childrens natl hosp	1	19	6
chitkara univ	2	1	6
csır cso	1	30	6
dıt univ	1	30	6
gulzar grp ınst	1	30	6
guru nanak dev engn coll	1	30	6
king khalid univ	1	30	6
manipal acad higher educ	1	30	6
univ cent punjab	1	19	6
univ educ	1	19	6
univ engn& technol taxila	1	19	6
air univ	1	40	5
chales Darwin univ	1	40	5
linnaeus univ	1	40	5
lovely profess univ	1	1	5
Middle east univ	1	1	5
princess nourah bint abdurrahmahman univ	1	1	5
symbiosis ınt deemed bello	1	40	5
univ andress bello	1	11	5
univ bernardo ohigging	1	11	5
univ cent chile	1	11	5
univ Santiago chile	1	11	5
vellore ınst technol	1	40	5
dongguk univ	1	237	4

Ortak Yazarlı Ülke Atıf Analizi (Co - authorship with Countries)

Ülke atıf analizi, konu ile ilgili yayınların ülkelere göre dağılımını ve aldıkları atıf sayıları arasındaki etkileşimi ortaya koymaktadır. Ülkenin en az 1 yayını ve en az 1 atıfı olsun. 39 ülkeden 32 ülke bu eşik değerini sağladığı ve 7 kümeden oluştuğu şekil 4'de görülmektedir. Atıf sayılarına dayalı sonuçlara bakıldığında, australia 4 yayın ve 483 atıfla ilk sırada, mauritius 1 yayın ve 440 atıfla ikinci sırada ve South korea 4 yayın ve 271 atıfla üçüncü sırada yer aldığı, makale sayılarına bakıldığında, ındia 13 yayın ve 217 atıfla ilk sırada, peoples r china 9 yayın ve 179 atıfla ikinci sırada ve saudi arabia 7 yayın ve 78 atıfla üçüncü sırada yer aldığı görülmektedir. Ülke atıf analizinden elde edilen etkileşim grafiği aşağıda Şekil 4'de sunulmuştur. En büyük halkaya sahip ülke en çok yayını olan ülkedir. Tablo 4'de aynı kümlerde yer alan ülkeler bulunmaktadır.

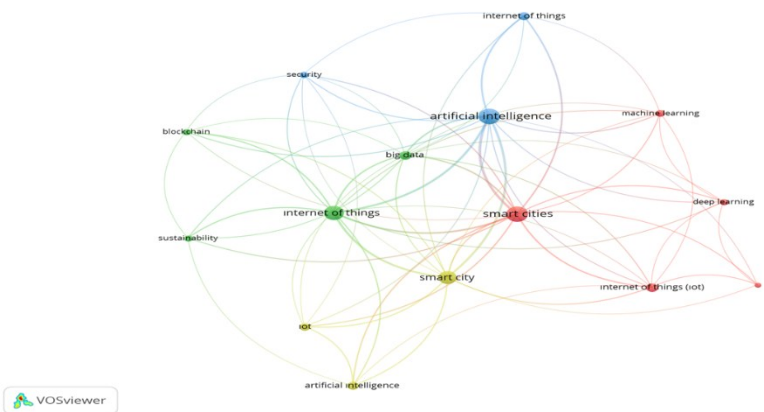


Şekil 4. Ülke Atıf Analizi ağ Haritası

Tablo 5. Ülke Atıf Analizi Ağ Kümesi

Küme1 Kırmızı	**Küme 2 Yeşil**	**Küme 3 Mavi**	**Küme 4 Açık yeşil**	**Küme 5 Mor**	**Küme 6 Mavi**	**Küme 7 Portakal**
Jordan Lebanon Pakistan sweden	Denmark India South Africa Thailand	Australia Mauritius singapore	Egypt Saudi arabia U arab emiratet	England Scotland South korea	Greece Peoples r chine	France Usa

Eş Anlamlı Kelime Analizi (Co-Occurrence with Author Keywords)



Şekil 5. Eş Anlamlı Kelime Analizi

Ortak kelime ağı, anahtar kelimelerin birlikte kullanım durumları doğrultusunda düğümleri ve bağlantıları göstermektedir. Renkleri aynı olan düğümler, kümeleri; düğüm boyutları ise kelime sıklıklarını ifade etmektedir. Beraber oluşma sıklığı, aynı kümedeki düğümlerin arasındaki mesafeden ve aralarındaki bağlantıların kalınlığından fark edilebilmektedir (Atabay vd., 2019; Hırlak, 2024).

Akıllı şehir, nesnelerin interneti, yapay zeka konusunda en çok kullanılan anahtar kelimelerin belirlenmesi ve bu değişkenler arasındaki ilişkilerin tespit edilmesi amacı ile anahtar kelime analizi yapıldı. En az 3 anaktar kelime bulunma koşulu baz alındığında, 171 anaktar kelimeden 15 anaktar kelime bu koşulu sağladı. Şekil 5'de birbiriyle ilgili kelimeler 4 küme oluşturmuştur. Tablo 6'da en çok tekrar eden kelimenin "artificial intelligence" olduğu görülmektedir.

Tablo 6. Eş Anlamlı Kelime Analizi küme dağılımı

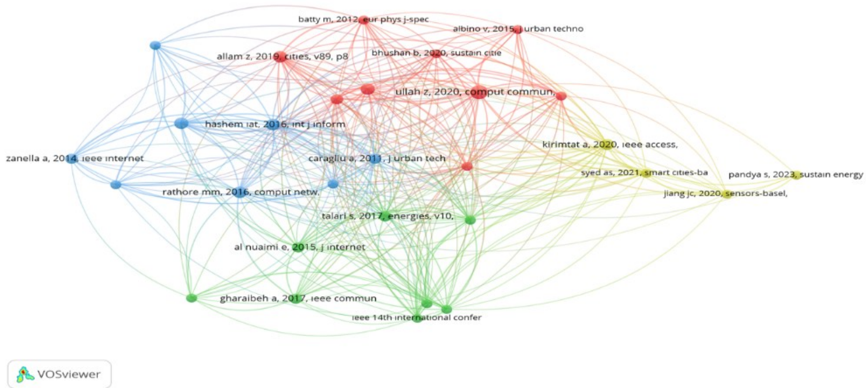
Anahtar Kelime	Tekrar Eden Anahtar Kelimeler	Etki gücü
artificial intelligence	25	64
smart cities	25	62
ınternet of things	22	54
smart city	19	42
big data	8	28
ınternet of things(ıot)	8	18
ınternet of things	7	16
Iot	6	16
artificial ıntelligence	7	15
Blockchain	4	13
Security	4	13
Sustainability	4	13
machine learning	5	12
deep learning	4	11
artificial intelligence(ar)	3	7

Ortak Atıf Analizi (Co-citation Analysis)

Ortak atıf analizi (Co-citation analysis), belirli yayınların ya da yazarların başka çalışmalarda birlikte ne sıklıkla atıf aldığını incelemektedir. Bu analiz, literatürdeki belirli araştırma alanlarını, temaları veya entelektüel yapıları ortaya çıkarmak için kullanılmaktadır.

Atıf Yapılan Referansların Ortak Atıf Analizi (Co-Citation with Cited References)

Orijinal örneklem 1375 yayından en az 4 kez atıf yapılan 29 yayına düşürüldü. Bu 29 en çok atıf yapılan referansa dayanarak, bu çalışma platform araştırma alanı içinde ortak atıf analizi yoluyla ağı oluşturdu. Sonuçlar 29 referansın dört kümeye ayrıldığını, her rengin bir kümeyi temsil ettiğini gösteriyor. Şekil 6 ’da atıf yapılan referansların ortak atıf analizi sonuçları görülmektedir.



Şekil 6. Atıf Yapılan Referansların Ortak Atıf Analizi Haritası (Co-Citation with Cited References)

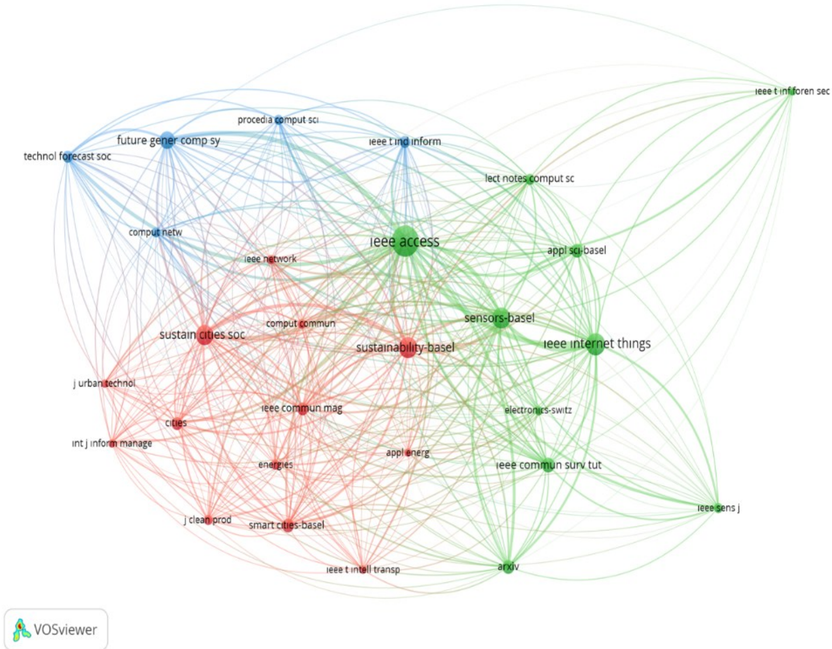
Tablo 7’de ilk on sırada yer alan yazarlardan, en yüksek ortak atıfa sahip yazarlar atıf sayısına göre sıralanmıştır.

Tablo 7. Atıf Yapılan Referansların Ortak Atıf Küme Analizi

Atıf Yapılan Kaynaklar	Atıf	Etki gücü
Hashem ıat, 2016, ınt j ınform manage, v36, p738, doı 1…	7	67
Ullah z, 2020, comput commun, v154, p313, doı10.1016	10	61
Kirimtat a, 2020, ıeee Access, v8, p86448, doı 10.1109/ac…	6	52
Allam z, 2019, cıtıes, v89, p80, doı 10.1016/j.cities.2019.0…	7	48
Caragliu a, 2011, j urban technol, v18, p65, doı 10.1080/…	5	45
Neirotti p, 2014, cıtıes, v38, p25, doı 10.1016/j.cities.2013…	7	45
Osman ams, 2019, future gener comp sy, v91, p620, doı…	5	45
Bhusman b, 2020, sustaın cıtıes soc, v61, doı 10.1016/j.s…	4	44
Xie jf, 2019, IEEE commun surv tut, v21, p2794, doı 10.1…	4	44
Silva bn, 2018, sustaın cıtıes soc, v38, p697, doı 10.1016/…	7	43

Atıf Yapılan Kaynağın Ortak Atıf Analizi (Co-Citation with Cited Sources)

Ortak atıf analizi, belirli bir araştırma alanında yürütülen çalışmalar arasındaki ilişkileri ve etkileşimleri inceleyen bir yöntemdir. Bu yöntem, bir makalenin referanslarını analiz ederek makalenin hangi diğer çalışmalara atıf yaptığını değerlendirir ve bu atıfların doğasını değerlendirir. Kaynak atıf analizi, konu hakkında en çok atıf alan yayınlar ile aldıkları atıf sayısı arasındaki ilişkiyi ortaya koyar. Kaynak atıf analizi yoluyla, 'Akıllı şehir', 'Nesnelerin Interneti' ve 'Yapay zeka' konularıyla ilgili, bir kaynağın en az atıf alma sayısı 15 olarak seçilerek 2220 kaynaktan 28 kaynak bu koşulu sağlayarak etkileşim grafiğinde üç bağlantı kümesi elde edilmiştir. Bu bağlamda, en çok atıf alan kaynağın 200 atıf ile 'IEEE access ' olduğu, bunu 104 atıf ile 'IEEE ınternet thıngs' izlediği görülmüştür. Şekil 7'de Atıf Yapılan Kaynağın Ortak Atıf Analizi sonuçları sonulmuştur.



Şekil 7. Atıf Yapılan Kaynağın Ortak Atıf Analizi Haritası

Tablo 8'de en yüksek atıf alan 14 kaynak sırasıyla gösterilmiştir.

Tablo 8. Atıf Yapılan Kaynağın Ortak Atıf Analizi

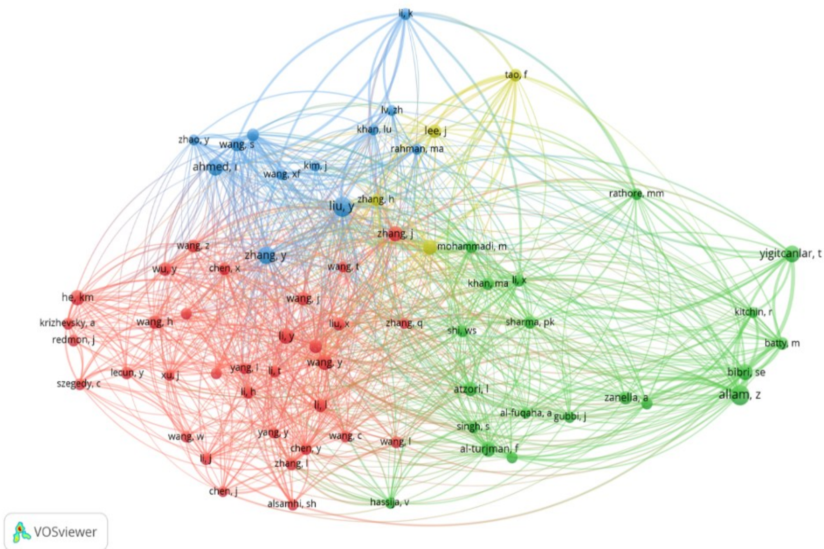
Kaynak	Atıflar	Kaynak	Atıflar
IEEE Access	200	lect notes comput sc	26
IEEE ınternet thıngs	104	j clean prod	23
sensors-basel	94	procedıa comput scı	22
sustaınabılıty-basel	89	comput netw	22
sustaın cıtıes soc	82	IEEE sens j	22
future gener comp sy	63	Energıes	21
IEEE commun surv tut	45	IEEE network	19
apply scı-basel	42	comput commun	18
Arxiv	39	electronıcs-swıtz	17
smart cıtıes-basel	38	ınt j ınform manage	16
IEEE commun mag	37	IEEE t ıntell transp	16
Cıtıes	35	IEEE t ınf foren sec	16
technol forecast soc	30	j urban technol	15
IEEE t ınd ınform	30	appl energ	15

Ortak Atıf Analizi (Co-Citation with Cited Authors)

Ortak atıf analizinde 4 kümede toplanma, akıl şehirlerde nesnelerin interneti, yapay zekâ konusunun 4 ana araştırma alanı veya tema olduğunu ve bu alanların her birinin belirli çalışmalar ve yazarlar etrafında şekillendiğini göstermektedir. Bu alanlar anahtar kelime analizi başlığı altında daha detaylı gösterilmektedir.

Bir yazarın en az 30 atıfı olsun. 34914 yazardan, 69 yazar bu eşik değerini sağladı. 69 yazar 4 küme oluşturdu. Bununla beraber bu ağların 4 farklı kümede toplandığı görülmüştür. Ortak atıf analizinde 4 kümede toplanma, 4 ana araştırma alanı veya tema olduğunu ve bu alanların her birinin belirli çalışmalar ve yazarlar etrafında şekillendiğini göstermektedir. Şekil 8'de en yüksek atıf alan iki yazar; Liu, y ve Allam, z, ortak atıf konusunda etrafında yoğunlaşılan yazarlar olarak görülmektedirler.

Analiz, aynı makalelerde birlikte görünen yazarların aldığı atıfları incelemeyi ve ortak yazarlar tarafından yürütülen işbirlikli araştırmanın etkisini değerlendirmeyi ve araştırma ağlarını belirlemeyi amaçlamıştır. Bu analizde, aynı makalede birlikte görünen yazarlara yapılan atıflar, aynı makaleye atıfta bulunan diğer makalelerde belirlenmiştir. Analizin sonuçları şekil 8'de aşağıda sunulmuştur.



Şekil 8. Yazarların ortak atıf analizinin dağılımı (Küme yoğunluklarının dağılımı)

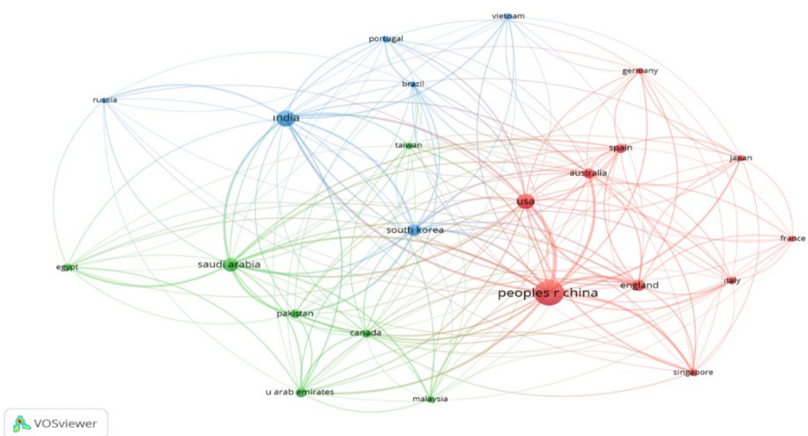
Tablo 9'de En yüksek atıf alan 10 yazar yer almaktadır.

Tablo 9. Yazarların Ortak Atıf Analizi

Yazar	Atıf
Liu, y	102
Allam, z	97
Zhang, y	76
Ahmed, ı	67
Yiğitcanlar, t	66
Wang, y	65
Li, y	63
Bibri, se	61
Zhang, j	57
Al-turjman, f	55

Ülke Atıf Analizi (Citation with Countries)

Ülke atıf analizi, konu ile ilgili yayınların ülkelere göre dağılımını ve aldıkları atıf sayıları arasındaki etkileşimi ortaya koymaktadır. Ülkenin en az yayın sayısı 21 ve en az atıf sayısı 20 olsun. 100 ülkeden 23 ülke bu koşulu sağladı. Atıf sayılarına dayalı sonuçlara bakıldığında, Peoples R China 430 yayın ve 10220 atıfla ilk sırada, USA 158 yayın ve 6111 atıfla ikinci sırada ve İngiltere 93 yayın ve 4123 atıfla üçüncü sırada yer aldığı görülmektedir. Ülke atıf analizinden elde edilen etkileşim grafiği şekil 9'da sunulmuştur.



Şekil 9. Ülkeye Göre Atıfların Ağ Görselleştirilmesi

Bilim alanında lider ülkeleri ve bu ulusların yıllar içindeki evrimini belirlemek için coğrafi bir analiz yapıldı. Ortak yazarlı ülkeler Tablo 10 'da sunulan bilimsel üretimin coğrafi dağılımına ilişkin verileri göstermektedir. Bibiriyle ilşkili ülkeler 3 kümeden oluşmaktadır. Tablo 11'de ülkelere göre yayın ve atıf sayısı yer almaktadır. Çin Halk Cumhuriyeti, USA, İngiltere, Hindistan şeklinde dağılım göstermektedir.

Tablo 10. Ülkeye Göre Atıfların Kümesi

Küme 1	Küme 2	Küme 3
Australia, England, France, Germany, ıtaly, japan, poeple r china, singapore, usa	Canada, egypt, malaysia, Pakistan, saudi arabia, taiwan, u arab emirates	Brazil, ındia, portugal, russia, South korea, vietnam

Tablo 11. Ülkelere Göre Küme Dağılımı

Ülke	Yayın	Atıf
People r China	430	10220
USA	158	6111
England	93	4123
Australia	67	3783
India	171	3019
Singapore	41	2756
Soulth Korea	95	2647
Saudi Arabia	129	2071
U Arap Emirates	65	1856
Canada	49	1385

Sonuç ve Değerlendirme

Bu çalışma literatür taramasından elde edilen çalışmaları özetler. WoS'ta elde edilen performans analizini kullanılarak, bu çalışma en çok atıf alan makaleler, en etkin yazarlar, en etkili dergileri, en etkili kurumları ve en çok yayın yapan ülkeleri en çok atıf alan referanslar belirlendi.

Bu çalışma, 2012-2024' yılları arasında web of science verilerini kullanarak, akıllı şehirler için literatürde yayınlanan araştırmalara genel bir bakış sağlamayı ve şu anda yürütülmekte olan araştırmanın kapsamlı bir görüntüsünü sunmayı amaçlamaktadır. Araştırmacılar, araştırmaya göre akıllı şehirler için Nesnelerin Interneti ve yapay zeka çalışmalarının 2016-2024 yılları arasında büyüme gösteren, hızla yükselen bir araştırma konusu olduğunu ortaya koymaktadır. Çin, Amerika Birleşik Devletleri, İngiltere Avustralya vs. olmak üzere ülkelerden çok çeşitli bilim insanları, özellikle Çin ve Amerika Birleşik Devletleri arasında uluslararsı iş birliği için zemin oluşturduğu görülmektedir. Bibliyometrik analiz kullanımıyla, araştırma ortamı, önemli makalelerin, kaynakların ve yazarların incelenmesi de dahil olmak üzere çeşitli düzeylerde araştırıldı. Yazarlar, nesnelerin interneti, yapay zeka, makine öğrenimi vs. akıllı şehirler üzerine yapılan araştırmaların akıllı şehirlerin sorunlarına yardımcı olmak için büyük bir potansiyele sahip olduğu sonucuna varıldı.

KAYNAKÇA

Atabay, E., Çizel, B., & Ajanovic, E. (2019). Akıllı şehir araştırmalarının R programı ile bibliometrik analizi. O. Emir (Ed.), Akıllı şehirler, 20. Ulusal Turizm Kongresi, Anadolu Üniversitesi Basımevi, 3, 1130-1137.

Hırlak, B. (2024). Yönetim alanındaki whıstleblowıng araştırmalarının bibliyometrik analizi. R&S – Research Studies Anatolia Journal, 7(2), 154-185. https://doi.org/10.33723/rs.1451312

Kurnaz, A. (2021). Etnosentrizm ile ilgili çalışmaların bibliyometrik analizi. Beykoz Akademi Dergisi, 9(2), 98-118. https://doi.org/10.14514/BYK.m.26515393.2021.9/2.98-118

van Eck, N.J., Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84,523–538. https://doi.org/10.1007/S11192-009-0146-3

Wang J., Jiang C., Zhang K., Quek T.Q., Ren Y., Hanzo L. (2018). Vehicular sensing networks in a smart city: Principles, technologies and applications. IEEE Wirel. Commun. 25:122–132. doi: 10.1109/MWC.2017.1600275.

Winkowska, J., Szpilko, D., & Pejić, S. (2019). Smart cityconcept in the light of the literature review. Engineer-ing Management in Production and Services, 11, 70-86. doi: 10.2478/emj-2019-0012 World's population increasingly urban with more than half living in urban areas | UN DESA | United Nations Department of Economic and Social Affairs. (2014). Retrieved April 23, 2017, from

http://www.un.org/en/development/desa/news/population/worldurbanization-prospects-2014.htm


BİDGE