ADVANCES IN BIOSTATISTICAL CLASSIFICATION: METHODS, TRENDS, AND MEDICAL APPLICATIONS

EDITOR

Assist. Prof. Dr. Elif ÜNAL ÇOKER

Prof. Dr. Öznur İŞÇİ GÜNERİ

Assist. Prof. Dr. Ecem DEMİR

Lecturer Dr. Burcu DURMUŞ

Assist. Prof. Dr. Aynur İNCEKIRIK

ISBN: 978-625-5753-15-1

Ankara -2025

ADVANCES IN BIOSTATISTICAL CLASSIFICATION: METHODS, TRENDS, AND MEDICAL APPLICATIONS

EDITOR

Assist. Prof. Dr. Elif ÜNAL ÇOKER
ORCID ID: 0000-0003-2572-3654

AUTHORS

Prof. Dr. Öznur İŞÇİ GÜNERİ[1]

Assist. Prof. Dr. Ecem DEMİR[2]

Lecturer Dr. Burcu DURMUŞ[3]

Assist. Prof. Dr. Aynur İNCEKIRIK[4]

[1]Muğla Sıtkı Koçman University, Faculty of Science, Department of Statistics, Kötekli Campus, Muğla, Türkiye
oznur.isci@mu.edu.tr
ORCID ID: 0000-0003-3677-7121

[2]Sivas Cumhuriyet University, Faculty of Science, Department of Statistics and Computer Science, Sivas, Turkey
ecemdemir@cumhuriyet.edu.tr
ORCID ID: 0000-0001-9714-0672

[3]Tekirdağ Namık Kemal University, Rectorate, Tekirdağ, Türkiye.
bdurmus@nku.edu.tr
ORCID: 0000-0002-0298-0802

[4]Manisa Celal Bayar University, Faculty of Economics and Administrative Sciences, Department of Econometrics, Manisa, Türkiye.
aynur.incekirik@cbu.edu.tr
ORCID ID: 0000-0002-5029-6036

DOI: https://doi.org/10.5281/zenodo.17669281

E mail: ubakyayinevi@gmail.com
www.ubakyayinevi.org
It is responsibility of the author to abide by the publishing ethics rules.
UBAK Publishing House – 2025©

ISBN: 978-625-5753-15-1
November / 2025
Ankara / Turkey

PREFACE

Recently, the volume of biological and clinical data has increased rapidly, which has made biostatistics even more of a vital field for advancing data-driven health research. Data structures are becoming more complex in various current areas of study. This trend arises due to the growth of greater amounts of data, the emergence of more powerful computing tools, and the widespread application of machine learning techniques. The recent changes have generated significant interest in the proper classification of information. This book presents three interrelated chapters. This work examines the evolution of classification techniques, the fundamental concepts underlying these methods, and their practical applications in the medical field via biostatistics.

The first chapter dives into classical and modern techniques of classification, including k-nearest neighbors, decision trees, logistic regression, and support vector machines. The chapter examines the methodological assumptions of each approach, demonstrating their strengths and weaknesses, and providing an objective assessment of their performance with various types of data. Contemporary research indicates an increasing integration of conventional methodologies with deep learning frameworks and ensemble techniques, resulting in hybrid models that enhance stability, precision, and interpretability. The chapter identifies several avenues for future research, including enhancing explainability, developing robust methods for small or imbalanced datasets, and creating integrated frameworks that combine statistical and machine learning concepts.

The second chapter presents a comprehensive bibliometric analysis of scientific literature concerning classification methods in biostatistics. This chapter provides an examination of 170 publications spanning almost four decades of research activity. The material illustrates primary trends in development, research clusters, patterns in international collaboration, and advancements in thematic areas. The results demonstrate a notable increase in research utilizing classification methods after 2015. The increase was mainly attributed to advancements in artificial intelligence, bioinformatics, and genomic data analysis, and the chapter further highlights a growing emphasis on classification studies designed for clinical application. This rise in number became especially evident during the COVID-19 pandemic, reflecting the multidimensional nature of modern biostatistical research.

The third chapter applies these methodological insights to an important public health problem: identifying cervical cancer behavioral risk factors using Support Vector Machines. Addressing class imbalance with SMOTE and evaluating several SVM kernels, the study demonstrates that polynomial-kernel SVM yields the most effective performance, particularly in modeling complex, nonlinear behavioral and psychosocial attributes. The findings demonstrate how machine learning-based models can significantly improve women's health through early diagnosis and preventive measures.

The chapters in this book collectively provide a comprehensive and prospective perspective on the integration of classification methods within biostatistics. They combine robust theory with practical

applications in contemporary healthcare systems. I sincerely hope this book becomes an essential resource for academics, physicians, data specialists, and students interested in the methodologies and practical aspects of biomedical classification. In summary, “Advances in Biostatistical Classification: Methods, Trends, and Medical Applications” is intended to be an important guide for the development of both ideas and real-world use in biostatistical classification.

21/11/2025

Assist. Prof. Dr. Elif ÜNAL ÇOKER

EDITOR

TABLE OF CONTENTS

CHAPTER 1

FOUNDATIONS AND ALGORITHMS OF CLASSIFICATION IN MACHINE LEARNING

Assist. Prof. Dr. Ecem DEMİR

INTRODUCTION

In today's data-driven world, classification has become one of the most widely used techniques in machine learning and statistical analysis. Many real-world problems, such as diagnosing a disease in the healthcare sector or determining whether an email is spam, can be automatically solved using classification algorithms. In recent years, the development of artificial intelligence-based methods has significantly increased the accuracy and generalizability of classification techniques [Goodfellow et al., 2016].

Classification falls under the category of supervised learning, and its purpose is to predict which class new observations belong to by learning from labeled examples in a given observation set. In this process, the advantages and limitations of different algorithms vary depending on the nature of the data set.

This section will examine both classical and modern classification methods in detail, starting with the basic principles of classification. It will also cover how to evaluate, compare, and improve the performance of algorithms.

1. CLASSIFICATION

Classification is the process of grouping objects based on their characteristics, allowing scientists to organize information into logically related categories for easier analysis and evaluation (Singh & Chauhan, 2012). The primary goal is to develop a model that can predict with the highest accuracy and in a generalizable manner which class an unlabeled new observation belongs to, based on examples with known labels during the training phase. In the context of data mining and machine learning, classification refers to learning a separation rule, decision surface, or probabilistic function from labeled (supervised) data to produce a mapping that can assign previously unseen examples to predefined classes (An, 2009; Kotsiantis, 2007). Therefore, classification is not only a descriptive but also a predictive task, and the success of the model is often evaluated using performance metrics such as accuracy, sensitivity/specificity, F1 score, ROC-AUC, or MCC for imbalanced datasets (James et al., 2021; Bishop, 2006).

Current literature demonstrates that classification has an extensive and interdisciplinary range of applications: clinical decision support and disease classification (e.g., cancer subtypes, diabetes complications), omics/data-intensive biomedical analysis, credit scoring and financial risk prediction, customer segmentation and churn analysis, network/cyber attack detection, fraud detection, image and speech recognition, and text/document classification are just some of these areas (Mlouhi & Hamdi, 2020; Han, Pei & Kamber, 2012; Aggarwal, 2015; James et al., 2021). Particularly in medical diagnosis

and bioinformatics applications, classification algorithms play a critical role in extracting meaningful patterns from multidimensional and noisy data, generating second opinions to support physician decisions, and creating risk scores for early diagnosis (Esteva et al., 2017; Chicco & Jurman, 2020). Similarly, in the field of cybersecurity, supervised classification methods enhance the accuracy of anomaly-based intrusion detection systems. In the field of text and natural language processing, they form the basis of tasks such as sentiment analysis, topic-based tagging, and multi-label document classification (Manning, Raghavan, & Schütze, 2008).

The application scope of classification remains very broad, encompassing early disease diagnosis, image and signal classification, network attack detection, credit risk scoring, churn prediction, and clinical phenotyping, among others (Esteva et al., 2021; Rajpurkar et al., 2022). Post-2020 literature reports that the hybrid use of deep learning-based classifiers with classical tabular methods (e.g., deep feature extraction combined with a tree-based classifier) yields meaningful gains in small and imbalanced clinical datasets (Huang et al., 2024).

Classification techniques are data mining methods used to separate and predict data samples into predefined classes or groups based on their features. Several studies highlight a few fundamental classification techniques (Rajwinder Kaur et al., 2017; Narsaiah Putta et al., 2018):

- Decision Trees: Hierarchical model for categorization

- k-Nearest Neighbor (k-NN): Classifies based on proximity to similar data points

- Support Vector Machines (SVM): Creates optimal separation boundaries between classes

- Artificial Neural Networks (ANN): Complex, brain-inspired computational models

These techniques are widely applied in various fields, including financial analysis, telecommunications, healthcare, and scientific research (Ms. Nalini Jagtap et al., 2017). Applications span various fields, including medical diagnosis, fraud detection, handwriting recognition, and drug discovery (An, 2009). The power of classification lies in its ability to process various types of data, predict group memberships, and support knowledge-based decision-making (Soofi et al., 2017).

2. CLASSIFICATION TECHNIQUES

2.1 k-Nearest Neighbor (k-NN)

The K-nearest neighbor (k-NN) algorithm is a distance-based classification and regression method. The method makes decisions based on the distance and similarity relationships between observations; in other words, it answers the question "Which class should an example

belong to?" by looking at the labels of the neighbors closest to that example in the feature space. For this purpose, data is represented as vectors in a multidimensional feature space, and each data point is positioned based on its distance from other points. Distances between data points are often calculated using metrics such as Euclidean, Manhattan, or cosine similarity; it is assumed that examples belonging to the same class are relatively closer to each other in this space (Alan, 2020; Cover & Hart, 1967). Thus, the class of a new observation is determined by the majority vote or weighted vote of its k nearest neighbors in the training dataset. The distance function is given as follows.

$$D = \left(\sum_{i=1}^{K} (|x_i - y_i|)^p\right)^{\frac{1}{p}}$$

Where

p=1; Manhattan distance

p=2; Euclidean distance

p=3; Minkowski distance.

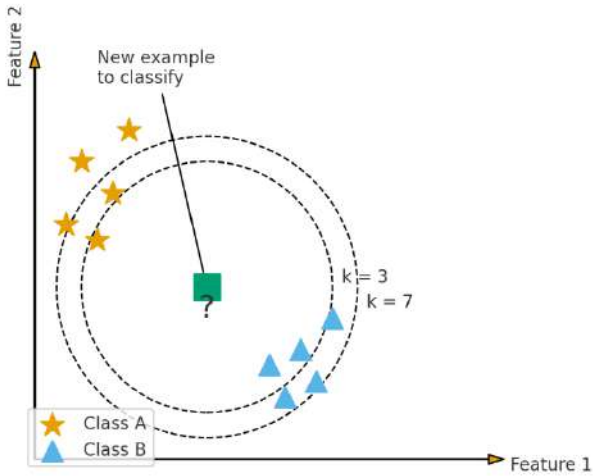

Figure 1. K-NN Algorithm

The k-NN algorithm graph is as shown in Figure 1. In the k-NN technique, the k parameter is a critical hyperparameter that determines the number of neighbors to be examined and directly affects the model's generalizability properties. Minimal k values can cause the model to be overly sensitive to noise and prone to overfitting. In contrast, tremendous k values can excessively smooth class boundaries, leading to the mixing of different classes. Therefore, the literature generally recommends selecting the k value specifically for the dataset using methods such as cross-validation (James et al., 2021). The simple, non-parametric structure of k-NN allows it to be easily applied to different types of datasets; however, because it is distance-based, it also makes preprocessing steps, such as feature scaling (normalization, standardization) and outlier control, critical.

The nearest neighbor approach can be categorized into two main types based on its structural assumptions: structured k-NN and unstructured k-NN (Wu et al., 2008; Bhatia, 2010). In unstructured k-NN, all training examples are stored in their raw form; for each new example, the distance to all these points is calculated, and the k points with the smallest distance are selected as the nearest neighbors. This "brute-force" approach is conceptually straightforward and yields exact results; however, as the data size and number of examples increase, the computational cost also increases, making it impractical for large-scale datasets [28].

In contrast, structure-based k-NN techniques focus on accelerating the search process by utilizing indexing and data organization mechanisms (e.g., k-d trees, ball trees, graph-based structures) that take into account the fundamental geometric structure of the dataset. In such structures, the training data is placed within a specific spatial or hierarchical structure; thus, for a new sample, the distance is calculated only on a limited number of regions or nodes that could be candidates, and the k-nearest neighbor search is significantly accelerated through this structure instead of performing a full scan across all data (Wu et al., 2008; Bhatia, 2010).

In conclusion, despite its simple distance-based principle, the k-NN algorithm is quite flexible and adaptable to different problem types when considered alongside decisions such as selecting the k parameter, determining the distance metric, and organizing the data structure. Structure-based approaches, in particular, contribute to the method's

applicability in high-dimensional and large-scale datasets by reducing the computational load of classical structureless k-NN.

2.2. Decision Tree Algorithms

The decision tree algorithm is a rule-based, hierarchical modeling approach used in both classification and regression problems within the scope of supervised learning. Decision trees are created in two steps: the first step is building the tree, and the second step is performing the classification. The basic idea is to divide the input space into more homogeneous subregions through successive binary (or multiple) splits and to obtain a relatively “pure” class distribution (or homogeneous numerical value) at each leaf node.

Thus, each path from the root node to the leaf node becomes a decision rule that can be expressed in the form of "if–then," and the model provides a decision process that can be easily followed by experts, serving not only as a statistical tool but also as a clear guide for informed decision-making. (Breiman et al., 1984; James et al., 2021).

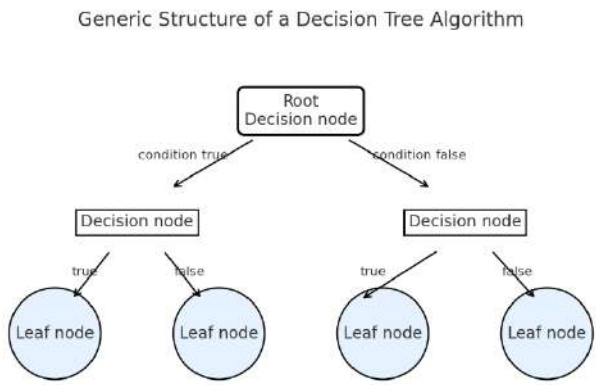

Figure 2. Decision Tree Algorithms

Decision trees have a hierarchical structure composed of nodes, branches, and terminal nodes, also known as leaves. The first node at the top level of the tree, where the decomposition process begins, is referred to as the root node. The endpoints where the decomposition process ends, no further splitting is performed, and the final class or output value is assigned are referred to as leaf nodes or, more commonly, pure nodes.

In the tree structure illustrated in the Figüre 2, although binary branching is performed from each decision node, multiple split branches from more than two nodes can also be designed, depending on the algorithm used and the problem structure. Therefore, decision trees are generally represented as binary structures, but they can also be generalized to include multi-way decision nodes when appropriate splitting criteria are defined.

In this type of decision tree representation, each node and leaf is more than just a schematic box; it also carries summary statistics related to the data set. Information such as the class distribution of the dependent variable, sample size, class ratios, and, when necessary, the error rate is typically included within the internal and leaf nodes. Thus, by looking at any point in the tree, one can quickly see the profile of the samples reaching that node in terms of the target variable. The branches indicate the value of the independent variable defining the split made at the relevant node, the category level, or the threshold range. For example, labels such as "Income > 5000 TL" or "Age ∈ [30, 45]" clearly show under which logical condition the data flow is directed to a sub-node. This structure enables the decision tree to be read both vertically (from root to leaf) and horizontally (between nodes at the same level), allowing for comparative interpretation.

One of the most important functions of decision trees is their ability to convert this visual structure into decision rules. Every path extending from the root node to a specific leaf node can be formulated as an "if–then" rule consisting of sequential conditions. For example, rules expressed as "If age>50 and blood pressure is high and cholesterol level>threshold, then Class = High Risk" are actually the textual equivalent of the branching order in the tree. These rules provide a rule base that can be used directly in expert systems, clinical decision support tools, or information systems where business rules are converted to automation, in addition to explaining how the model works.

When performing branching in a decision tree, deciding which independent variable to split on is a critical decision. The criteria used to make this selection, along with their corresponding mathematical expressions, are presented below.

Entropy is a measure that quantitatively expresses the level of uncertainty or disorder contained in a random variable or class distribution (Altunkaynak, 2017). In other words, it defines the degree of unpredictability in the system. If all observations of a variable are concentrated in a single value or a single class, i.e., if the variable has an entirely homogeneous structure, uncertainty is negligible and the entropy value is at a minimum. Conversely, if the possible values or classes of the variable are observed with approximately equal probability, the disorder and unpredictability in the system increase; in this case, entropy reaches its maximum value (Cover & Thomas, 2006; MacKay, 2003).

When a random Y variable has k different levels (classes), the entropy associated with this variable can be defined as follows:

$$H(Y) = H(p_1, p_2, \dots, p_k) = \sum_{j=1}^{k} \left(p_j \log_b \left({}^{1}/_{p_j} \right) \right)$$

Here, p_j represents the probability of occurrence of level j (class) of the variable Y. b is the base of the logarithm. When the variable has two levels ($k=2$), the base of the logarithm is typically taken as $b=2$, and the resulting measure is known as Shannon entropy.

When the variable has more than two categories, *b=10* is used, and Hartley entropy is employed.

In the context of classification and decision trees, entropy is used to measure the degree of "mixedness" in the class distribution at a node. If all examples at a node belong to a single class, there is no uncertainty; in this case, the entropy approaches zero, and the node is considered "pure." Conversely, if the examples in a node are distributed approximately equally among different classes, it becomes difficult to predict which class they belong to; in this case, entropy reaches its maximum value (Bishop, 2006; James et al., 2021). Therefore, in decision trees, when branching decisions are made, the attribute and threshold values that reduce entropy the most (i.e., reduce uncertainty the most) are preferred; the information gain measure is also directly based on this principle (Han, Pei, & Kamber, 2012; Aggarwal, 2015).

Gain, is a measure that quantitatively expresses how much "information" a split adds or how much uncertainty it reduces, particularly in the context of decision trees. In other words, when we branch a node based on a specific independent variable (feature), it measures whether this branching makes the class structure of the dependent variable more regular (more pure). For a categorical X_i independent variable, information gain can be defined as follows:

$$Gain(X_i) = H(Y) - \sum_{j=1}^{k_i} P(X_{ij}) H(Y|X_{ij}); \qquad i = 1,2,\dots,m$$

H(Y): It shows the initial entropy value of the dependent variable Y, that is, the level of uncertainty before any division is made.

k_i: Number of categories of the independent variable X_i

$P(X_{ij})$: It is the probability of occurrence of level j of the independent variable X_i, and therefore represents the weight of the relevant subgroup within the entire data set.

$H(Y|X_{ij})$: When the independent variable X_i is at level j, that is, under the condition, $X_i = X_{ij}$, it is the conditional entropy value of the dependent variable Y; in other words, it measures the level of uncertainty regarding the class distribution in this subgroup.

m: It shows the total number of independent variables in the model.

An independent variable with high information gain divides the data set into more homogeneous subsets in terms of classes, thereby further reducing uncertainty. Therefore, when selecting the feature to branch on in decision tree algorithms, the variable with the highest $Gain(X_i)$ value is preferred.

2.3. Logistic Regression

Logistic regression is a probabilistic and parametric model used for classification problems. In its most common form, binary logistic regression, the dependent variable Y has two categories (0/1), where 1

indicates the occurrence of the event and 0 indicates its non-occurrence. The primary objective is to develop a model that best explains the behavior of the dependent variable using the smallest possible vector of independent variables, X, and produces the most accurate predictions for future observations, thereby determining the probability that an observation belongs to a specific class (Y=1). (Hosmer, Lemeshow & Sturdivant, 2013; James et al., 2021; Alan and Karabatak, 2020).

In some cases, the researcher can control the levels of independent variables through experimental design. In applications where this is possible, having at least 30 observations in each "cell" (group) corresponding to the levels of (X_i) significantly increases the model's fit to the data and the reliability of the results due to large sample properties (asymptotic convergence, normal approximation) (Bircan, 2004).

Binary logistic regression models the probability $P(Y = 1|X)$ using the logit link function, rather than the class label directly:

$$\pi(X) = P(Y = 1|X),$$

$$logit\big(\pi(X)\big) = log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_{1x_1} + \cdots + \beta_{px_p},$$

From here,

$$\pi(X) = \frac{1}{1 + exp\left(-\left(\beta_0 + \beta_{1x_1} + \cdots + \beta_{px_p}\right)\right)}$$

The expression is obtained. Thus, the linear combination of explanatory variables is defined in terms of the log-odds, and the output probability is modeled in the 0-1 range (Hastie, Tibshirani, & Friedman, 2009).

One of the most important features of logistic regression is that the coefficients can be interpreted in terms of odds ratios. The β_j coefficient represents the marginal effect of a one-unit increase in x_j on the log-odds, holding other variables constant; $\exp(\beta_j)$ represents the multiplier effect of the same increase on the odds ratio. For example, if $\exp(\beta_j) = 1{,}5$, a one-unit increase in x_j increases the odds of the event occurring by a factor of 1.5. This feature makes logistic regression particularly suitable for applications in medical risk factor analysis, epidemiology, and the social sciences (Kleinbaum & Klein, 2010; Menard, 2010).

2.4. Support Vector Machines (SVM)

Support vector machines (SVM) are a family of methods explicitly developed for binary classification problems within the scope of supervised learning. They have a strong theoretical foundation and are widely used in practice (Schölkopt & Smola, 2002; Jabardi, 2025).

SVMs are considered one of the most fundamental yet theoretically advanced classification approaches used in machine learning. Compared to neural network-based models, SVMs can deliver stable results even with relatively small sample sizes and, due to their convex

optimization-based structure, are less prone to overfitting (Jabardi, 2025).

Within the SVM framework, each observation is represented as a point in an N-dimensional feature space; the coordinates of these points indicate the feature values of the corresponding unit. The classification process is performed by defining a hyperplane in this feature space. A hyperplane corresponds to a line in two-dimensional space, a plane in three-dimensional space, and a generalized version of this concept in higher dimensions. The goal is to obtain a separating surface where all points belonging to one class remain on one side of the hyperplane and all points belonging to the other class remain on the other side (Sarker, 2021; Prasad et al., 2023).

Suppose multiple hyperplanes can separate the same dataset. In that case, the SVM attempts to select the hyperplane that best separates the classes, i.e., the one that maximizes the distance between the hyperplane and the closest points belonging to both classes. This minimum distance is referred to as the "margin." The points closest to the separating hyperplane that define the margin are called support vectors and play a critical role in determining the model's decision boundary (Jabardi, 2025). For a SVM to learn a separating hyperplane, it requires a training dataset where each observation belongs to a predefined class and is correctly labeled. Therefore, SVM falls under the class of supervised learning algorithms that operate based on input-output mapping.

The method solves a convex optimization problem in the background that maximizes the margin between classes and ensures that the points belonging to each class remain as close as possible to the "correct" side of the hyperplane. This convex structure supports the conclusion that the obtained solution is a global optimum and that the model exhibits statistically good generalization properties (Otchere et al., 2021; Zulfiqar et al., 2022). Although SVMs are fundamentally designed for binary classification problems, in practice, multi-class situations can also be handled through various strategies. The most common approaches are one-vs-all, which separates each class from all others, and one-vs-one, where a separate binary classifier is trained for each class pair (Alwahedi et al., 2024). Thanks to such schemes, SVM can be effectively used in multi-class classification problems as well.

Maximum margin hyperplane

In the context of binary classification, each observation is represented by $x \in \mathbb{R}^N$ and class labels

$$y_i \in \{-1, +1\}, \quad i = 1,2, \dots, n$$

SVM defines a hyperplane that separates these observations:

$$w^T x + b = 0$$

Where:

w: Weight (parameter) vector,

b: Bias term.

The classification rule is as follows:

$$\hat{y}(x) = sign(w^T x + b)$$

That is, $w^T x + b > 0$, the class is predicted as +1, and if $w^T x + b < 0$, the class is predicted as -1. The margin is expressed as the distance between the hyperplane and the nearest points (support vectors). $\|w\|$ is the Euclidean norm of the weight vector w.

$$Marj = \frac{2}{\|w\|}$$

SVM minimizes $\|w\|$ in order to maximize this margin. A graphical representation of the SVM algorithm is provided in Figure 3 (Jacardi, 2025).

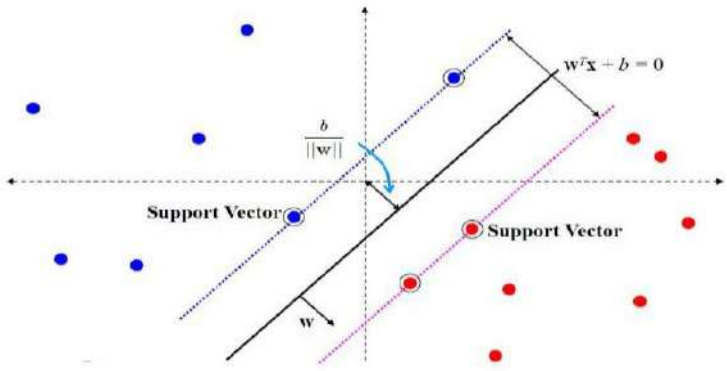

Figure 3. Support Vector Machines

3. DISCUSSION

The k-nearest neighbors, decision trees, logistic regression, and support vector machines discussed in this section are among the most

fundamental and frequently used methods in the field of supervised classification. The fact that each algorithm has different assumptions, data requirements, and computational costs highlights the importance of selecting methods based on problem characteristics and/or using hybrid structures that combine methods, rather than a "single best method" approach in real-world applications. Recent studies have shown that, particularly in critical areas such as healthcare, energy, and cybersecurity, classical classifiers are often used in conjunction with deep learning or ensemble models, thereby producing balanced solutions in terms of both accuracy and interpretability (Esteva et al., 2021; Rajpurkar et al., 2022).

k-NN offers competitive performance in small and medium-sized, low-dimensional datasets due to its non-parametric and heuristic structure; however, it is known that computational costs increase rapidly as the number and size of examples increase. Therefore, recent studies report that k-NN is used in conjunction with data structure-sensitive indexing techniques (k-d trees, ball-trees, graph-based structures) or dimension reduction methods, thereby reducing both search time and noise sensitivity (Kulkami & Babu, 2013; KR et al., 2025). Cui et al. (2003) provide a concrete example using a Δ-tree that employs Principal Component Analysis to reduce dimensionality, enabling more efficient search by pruning search areas and reducing distance calculation costs. These techniques collectively address fundamental challenges of k-NN, such as computational complexity and high sensitivity to high-dimensional noise.

While decision trees are highly interpretable, the high variability of individual tree models and their tendency toward overfitting have led to the extension of these methods with tree-based ensemble approaches, such as Random Forest, XGBoost, LightGBM, and CatBoost. These models have been shown to provide meaningful performance gains, particularly on complex, heterogeneous, and imbalanced datasets (Breiman, 2001; Chen & Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018; Lundberg et al., 2020).

Logistic regression remains the reference method, particularly in medicine and the social sciences, where risk factors need to be interpreted quantitatively, thanks to its linear log-odds assumption and robust statistical foundation (Hosmer et al., 2013; James et al., 2021). However, post-2020 literature shows that logistic regression models with L1/L2 or elastic-net regularization applied to high-dimensional and multicollinear datasets demonstrate superior performance compared to classical models in terms of both variable selection and generalizability (Friggeirsson et al., 2024; El Guide et al., 2022).

Support vector machines, on the other hand, offer a powerful alternative for problems involving high-dimensional and nonlinear decision boundaries, thanks to the maximum margin principle and kernel functions; its flexibility is highlighted in studies such as oil reservoir property estimation (Otchere et al., 2021), electricity load forecasting (Zulfiqar et al., 2022), and robust SVM variants (Prasad et al., 2023).

Support vector machines provide a powerful alternative for problems involving high-dimensional and nonlinear decision boundaries, thanks to the maximum margin principle and kernel functions. They have demonstrated superior performance compared to traditional neural networks in petroleum engineering, achieving outstanding success in reservoir property estimation (Otchere et al., 2021). In electricity load prediction, SVMs effectively model complex nonlinear relationships and provide accurate predictions by incorporating multiple input factors (Türkay et al., 2011; Acera, 2010). Their robustness stems from not assuming prior data distribution and effectively processing high-dimensional, complex datasets (Prasad et al., 2023; Van Messem, 2020).

On the other hand, explainable AI discussions, particularly in regulated fields such as healthcare and finance, demonstrate that not only prediction accuracy but also the transparency and interpretability level of model decisions are at least as important as accuracy. The combined use of explainability techniques developed for tree-based ensemble models, such as SHAP and similar methods (Lundberg et al., 2020), with statistical interpretability-rich methods like logistic regression and decision tree types, is a prominent trend of recent times (Molnar, 2022; Huang et al., 2024). In this context, the classical classification algorithms discussed in this section remain an important fundamental reference and the first set of methods to be considered in most applications, alongside modern deep learning and ensemble approaches.

4. CONCLUSION and FUTURE WORK

This section discusses the theoretical framework and fundamental components of classification problems, including KNN, decision trees, logistic regression, and support vector machines, which are discussed in detail.

The strengths and weaknesses of each method, model assumptions, and application areas are evaluated comparatively in light of the current literature. In general, the non-parametric and straightforward nature of k-NN, the rule-based and interpretable structure of decision trees, the probabilistic and statistically rich framework of logistic regression, and the strong generalization capacity of SVM based on maximum margin and kernel methods make these methods indispensable for both educational purposes and real-world applications. However, current data issues such as large and high-dimensional datasets, class imbalance, missing observations, and label noise indicate that these algorithms require careful preprocessing, appropriate model selection, and hyperparameter tuning rather than direct and "out-of-the-box" application (James et al., 2021).

Post-2020 studies reveal that classical classification methods are increasingly being used as components of hybrid and ensemble structures. For example, combining representations learned with deep neural networks (deep features) with tree-based classifiers or SVM provides meaningful performance improvements, especially in small and imbalanced medical datasets (Esteva et al., 2021; Huang et al.,

2024). Similarly, the use of ensemble learning approaches (Random Forest, XGBoost, LightGBM, CatBoost) alongside base classifiers in fields such as energy demand, financial risk, and engineering processes enables the development of more stable and generalizable models in complex and noisy data structures (Breiman, 2001; Chen & Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018; Zulfiqar et al., 2022).

Future work is likely to focus on three main areas:

(i) Developing methods that enhance the explainability of classification algorithms and ensuring these methods comply with regulatory requirements;

(ii) Designing robust classification strategies with high sample efficiency for small, imbalanced, and high-dimensional datasets;

(iii) Practical testing of hybrid frameworks that integrate classical statistical models with deep learning architectures.

In this context, both the theoretical foundations and practical advantages of the methods presented in this section provide a solid foundation for the development of complex models in the future and contribute to shaping the research agenda in the field of classification.

REFERENCES

Acera, M. M. M. (2010). Electricity Load Forecasting Using Machine Learning Techniques. In Business Intelligence in Economic Forecasting: Technologies and Techniques, pp. 318-336. IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-60960-818-7.ch313.

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.

Alan, A., & Karabatak, M. (2020). Veri seti-sınıflandırma ilişkisinde performansa etki eden faktörlerin değerlendirilmesi. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 32(2), pp. 531-540. https://doi.org/10.35234/fumbd.738007.

Altunkaynak B. (2017). Veri Madenciliği Yöntemleri ve R Uygulamaları. Seçkin yayınevi, Ankara.

Alwahedi F, Aldhaheri A, Ferrag MA, Battah A, Tihanyi N. (2024). Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models. Internet of Things and Cyber-Physical Systems. 4, pp. 167-185. https://doi.org/10.1016/j.iotcps.2023.12.003.

An, A. (2005). Classification Methods. In J. Wang (Ed.), Encyclopedia of Data Warehousing and Mining, pp. 144-149. IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-59140-557-3.ch028.

Bhatia N. (2010). Survey of nearest neighbor techniques. arXiv preprint arXiv:1007.0085.

Bircan, H. (2004). Lojistik regresyon analizi: Tıp verileri üzerine bir uygulama. Kocaeli Üniversitesi Sosyal Bilimler Dergisi, 8, pp. 185-208.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Wadsworth.

Breiman, L. (2001). Random forests. Machine Learning, *45*(1), pp. 5–32. https://doi.org/10.1023/A:1010933404324.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794. https://doi.org/10.1145/2939672.2939785.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC). *BMC Genomics*, 21(6).

Cover T., Hart P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory 1967. 13, pp. 21-27. https://doi.org/10.1109/TIT.1967.1053964.

Cui, B., Ooi, B. C., Su, J., & Tan, K. L. (2003). Contorting high dimensional data for efficient main memory KNN processing. In Proceedings of the 2003 ACM SIGMOD international conference on Management of data, pp. 479-490. https://doi.org/10.1145/872757.87281.

Esteva, A. et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 542, pp. 115–118.

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y. et al. (2021). Deep learning-enabled medical computer vision. npj Digital Medicine. 4(1), 5. https://doi.org/10.1038/s41746-020-00376-2.

El Guide, M., Jbilou, K., Koukouvinos, C., & Lappa, A. (2022). Comparative study of L_1 regularized logistic regression methods for variable selection. Communications in Statistics-Simulation and Computation. 51(9), pp. 4957-4972.

https://doi.org/10.1080/03610918.2020.1752379.

Fridgeirsson, E. A., Williams, R., Rijnbeek, P., Suchard, M. A., & Reps, J. M. (2024). Comparing penalization methods for linear models on extensive observational health data. Journal of the American Medical Informatics Association. 31(7), pp. 1514-1521. https://doi.org/ 10.1093/jamia/ocae109.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Han, J., Pei, J., & Kamber, M. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

Huang, Y. et al. (2024). Hybrid deep feature extraction and tree-based classification for small-scale medical datasets. IEEE Journal of Biomedical and Health Informatics, 28(2). https://doi.org/10.1109/ACCESS.2023.3304628.

Jabardi, M. (2025). Support Vector Machines: Theory, Algorithms, and Applications. Infocommunications Journal, 17(1). https://doi.org/10.36244/ICJ.2025.1.8.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning with applications in R (2nd ed.). Springer.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient

boosting decision tree. In Advances in neural information processing systems, 30.

Kleinbaum, D. G., & Klein, M. (2010). Logistic regression: A self-learning text (3rd ed.). Springer.

KR, M., Kurban, H., Kulekci, O. M., & Dalkilic, M. M. (2025). Telescope indexing for k-nearest neighbor search algorithms over high-dimensional data & large data sets. Scientific Reports, *15*(1), 24788. https://doi.org/10.1038/s41598-025-09856-5.

Kotsiantis, S. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31, pp. 249–268.

Kulkarni, S. G., & Babu, M. V. (2013). Introspection of various K-nearest neighbor techniques. UACEE International Journal of Advances in Computer Science and Its Applications, 3(2), pp. 103-106.

Lundberg, S. M., Erion, G., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, *2*(1), pp. 56–67.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge.

Menard, S. (2010). Logistic regression: From introductory to advanced concepts and applications. Sage.

Mlouhi, Y., & Hamdi, M. A. (2020). Statistical Analysis and Segmentation IVUS Images. In 2020, the 4th International Conference on Advanced Systems and Emergent Technologies (IC_ASET), pp. 253-256. https://doi.org/10.1109/IC_ASET49463.2020.9318252.

Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (Updated ed.). Lulu.

Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. Oriental Journal of Computer Science and Technology, *8*(1), pp. 13-19.

Otchere, D. A., Ganat, T. O. A., Gholami, R., & Ridha, S. (2021). Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. Journal of Petroleum Science and Engineering, 200, 108182. https://doi.org/10.1016/j.petrol.2020.108182.

Prasad, S.C., Anagha, P. & Balasundaram, S. (2023). Robust Pinball Twin Bounded Support Vector Machine for Data Classification. Neural Process Lett, 55, pp. 1131–1153. https://doi.org/10.1007/s11063-022-10930-6.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical

features. Advances in Neural Information Processing Systems, 31. https://doi.org/10.48550/arXiv.1706.09516.

Rajpurkar P, Chen E, Banerjee O, Topol EJ. (2022). AI in health and medicine. Nature Medicine;28(1), pp. 31-38. https://doi.org/10.1038/s41591-021-01614-0.

Sarker, I.H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160. https://doi.org/10.1007/s42979-021-00592-x.

Singh, M., & Chauhan, B. (2012). Classification: A holistic view. International Journal for computer science and communication, *3*(1), pp. 69-72.

Soofi, A. A., & Awan, A. (2017). Classification techniques in machine learning: applications and issues. Journal of Basic & Applied Sciences, 13, pp. 459-465. https://doi.org/10.6000/1927-5129.2017.13.76.

Türkay, B. E., & Demren, D. (2011). Electrical load forecasting using support vector machines. In 2011 7th International Conference on Electrical and Electronics Engineering (ELECO), Bursa, Turkey, pp. 49-53.

Van Messem, A. (2020). Support vector machines: A robust prediction method with applications in bioinformatics. In Handbook of

statistics, 43, pp. 391-466. Elsevier. https://doi.org/10.1016/bs.host.2019.08.003.

Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, et al. (2008.) Top 10 algorithms in data mining. Knowledge and Information Systems; 14, pp. 1-37. https://doi.org/10.1007/s10115-007-0114-2.

Zulfiqar M, Kamran M, Rasheed MB, Alquthami T, Milyani AH. (2022). Hyperparameter optimization of support vector machine using adaptive differential evolution for electricity load forecasting. Energy Reports, 8, pp. 13333-13352. https://doi.org/10.1016/j.egyr.2022.09.188.

CHAPTER 2

DEVELOPMENT OF CLASSIFICATION APPROACHES IN BIOSTATISTICS: A BIBLIOMETRIC REVIEW BASED ON WEB OF SCIENCE

Assist. Prof. Dr. Ecem DEMİR

INTRODUCTION

Biostatistics is a science that deals with the collection, analysis, interpretation, and presentation of biological and medical data. It plays an important role in the development of decision support systems and diagnostic models in modern medicine.

Classification problems are supervised learning approaches that aim to classify data into predefined categories. Classification methods, as one of the basic building blocks of statistical learning, are of increasing importance, especially in the field of biostatistics. The generation of health data in increasingly larger volumes and more complex structures necessitates the use of practical classification algorithms on these data. Classification algorithms play a crucial role in the analytical support of clinical processes, including diagnosis, treatment decisions, and risk assessment. In this context, classification-based scientific production in the field of biostatistics is increasing, allowing for multidisciplinary applications. This book chapter aims to examine the directions of scientific production, collaboration networks,

thematic clusters, and research trends by analyzing publications in biostatistics with a bibliometric approach in the context of classification methods. Especially in recent years, the increasing use of machine learning and artificial intelligence algorithms with health data has increased the interest in classification algorithms (Kourou et al., 2015).

In this chapter, the developmental trends, production volumes, and research focuses of scientific publications on classification methods in the field of biostatistics are analyzed using bibliometric methods. Classification algorithms form the basis of critical decision support systems, such as diagnosis, risk stratification, and prediction, in the analysis of biomedical data. In recent years, academic interest in these methods has increased rapidly, accompanied by a significant rise in the number of publications. This increase is associated with both advances in computational technologies and the growth in the volume of biological and clinical data. This book chapter aims to guide researchers by analyzing the structural and contextual characteristics of publications in this field.

1. CONCEPTUAL FRAMEWORK

1.1. Definition and Importance of Classification Methods

Classification is a statistical process that aims to categorize samples in a dataset into specific groups or classes. These methods enable individuals or samples to be classified according to a specific outcome (e.g., presence or absence of a disease). Classification is a crucial data mining technique used to categorize items into predefined

classes or groups based on their characteristics (Kesavaraj & Sukumaran, 2013; Archana & Elangovan, 2014). Among the standard algorithms, there are various classification methods such as decision trees, neural networks, support vector machines, k-nearest neighbors, and Naive Bayes, which are applied in various fields such as text classification, healthcare, and image recognition (Archana & Elangovan, 2014; Kaur & Verma, 2017; Sabouri et. al., 2022). The main goal of these methods is to develop predictive models to facilitate decision-making in multivariate data environments. These techniques are applied in various industries to identify and group data efficiently (Gupta & Aggarwal, 2010). In image classification, specialized techniques such as, Minimum Distance, Maximum Likelihood, Artificial Neural Networks, and Support Vector Machines are used to extract information from digital images (Thakur & Maheshwari, 2017). Classification algorithms have different advantages and disadvantages that researchers analyze to determine their suitability for specific applications (Khujaev et. al. 2023).

Researchers analyze these algorithms based on criteria such as accuracy, speed, efficiency, and scalability to determine their suitability for different tasks (Sabouri et al., 2022). Algorithm selection depends on the specific problem and dataset characteristics, as there is no universal method that works best for all scenarios (Kalcheva et al., 2020). Challenges in classification include model reliability and performance evaluation. Techniques such as K-Fold Cross-Validation have been proposed to facilitate more accurate evaluations (Khujaev et

al., 2023). As the field evolves, researchers continue to develop new algorithms and improve existing algorithms to address existing challenges in classification (Fan-Zi & Qiu, 2004).

In general, classification techniques play a vital role in transforming large datasets into understandable and actionable information. The primary goal of these methods is to develop predictive models that facilitate informed decision-making in multivariate data environments. Especially in the health sciences, classification techniques are widely used for early disease diagnosis, determining individual risk levels, and creating personalized medical approaches.

1.2. Areas of Use in Biostatistics

Classification techniques play a crucial role in biostatistics and healthcare applications. These methods are used for disease diagnosis, predicting patient outcomes, and identifying risk factors (Goel & Kumar, 2023). Various algorithms, including decision trees, logistic regression, support vector machines, and neural networks, are used in medical data analysis (Khan et al., 2020). The performance of these classifiers is evaluated using statistical metrics and significance tests, with caution advised when interpreting results from imbalanced datasets (Wang et al., 2018).

Classification techniques have evolved from traditional statistical methods to more advanced machine learning approaches, enabling the handling of the increasing volume of biological and medical data (Fielding, 2006). Applications extend to image recognition in radiology

and pathology (Goel & Kumar, 2023). Ensemble methods, such as boosting, bagging, and stacking, are also employed in healthcare decision-making systems (Khan et al., 2020). The effectiveness of classification methods in healthcare applications has been demonstrated across various medical conditions, including thyroid, cancer, heart disease, and diabetes (Jha et al., 2018). Overall, classification techniques enable healthcare professionals to make more informed decisions based on patient data (Goel & Kumar, 2023).

2. MATERIALS AND METHODS

In this study, the WoS Core Collection database was used to search for articles published between 18 July 2025, starting from the first publication on classification techniques in biostatistics in 1988. Study data were obtained by Boolean search using keywords (TS='Classification' AND TS='Biostatistics'). Although there were 177 publications in total, only 170 research articles, book chapters, reviews, and proceedings were included in the analysis. Duplicate records and irrelevant studies were excluded during the data cleaning and pre-processing stages.

The open-source R-based Bibliometrix package (Aria & Cuccurullo, 2017) and VOSviewer software (Van Eck & Waltman, 2010) were used for data analysis and visualization. In addition, descriptive statistics, co-authorship networks, and keyword co-occurrence analysis were performed for thematic analysis and inference of collaboration networks.

RESULTS

Descriptive analyses were conducted utilizing Biblioshiny. The primary data insights are depicted in Figure 1.

Figure 1. Main Information

This bibliometric overview provides a quantitative snapshot of the scholarly landscape within the specified research domain between 1988 and 2025. The dataset comprises a total of 170 documents contributed by 1,017 authors, reflecting a collaborative and expanding body of literature. The timespan from 1988 to 2025 encompasses nearly four decades of academic output. An annual growth rate of 5.4% suggests a steady and positive increase in publication activity over time, indicating growing interest and scholarly engagement in the field. Publications are distributed across 118 different sources (e.g., journals or conference proceedings), evidencing a moderately diverse dissemination of research. The relatively compact volume of 170 documents implies a focused but active area of investigation. With only 10 documents authored by a single researcher, the field is highly collaborative, as also supported by the average of 6.57 co-authors per document. Notably, 28.24% of contributions involve international

collaboration, underscoring the global nature of research efforts in this area. The presence of 666 unique author keywords (DE: Author's Keywords) reveals a broad and evolving conceptual scope, suggesting multidimensional thematic diversity.

An impressive average of 75.89 citations per document reflects the high scholarly impact of publications within this field. A total of 5,961 references across 170 documents indicates deep engagement with the literature. In contrast, the average document age of 9.64 years suggests that the field maintains relevance through both historical and contemporary studies.

2.1 The Annual Publication Distribution Map Index:

The annual scientific production chart illustrates the annual volume of scientific publications over a 37-year period. The data reveals several distinct phases in the evolution of scholarly activity:

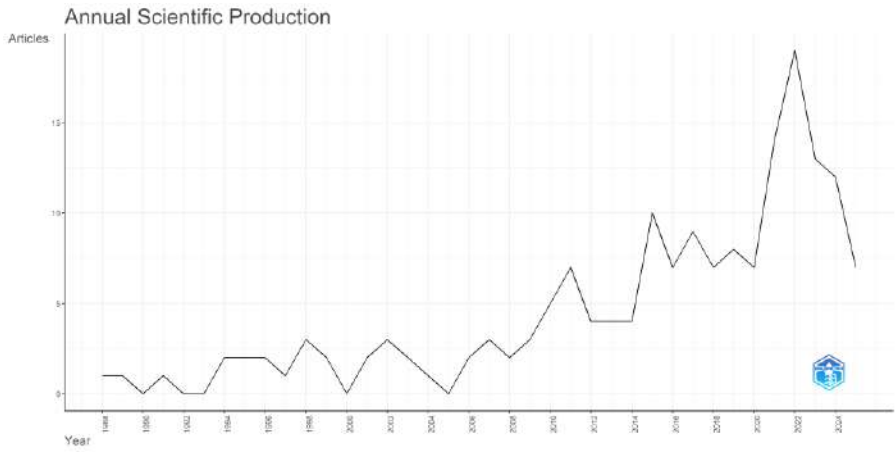

Figure 2. The Annual Number of Scientific Production

During the early decades, publication rates remained modest, typically fewer than five articles per year. This period likely represents the foundational phase of the field, characterized by limited but pioneering contributions. A gradual increase in annual output is observed, with intermittent fluctuations. This phase marks the emergence of growing scholarly interest and the establishment of the field as a distinct research area.

Scientific production surged sharply, peaking around 2021 with over 17 articles published in a single year. This likely corresponds to intensified research activity driven by technological advances, funding influxes, or societal relevance. A moderate decline in output has been noted in recent years. While this may reflect stabilization or shifts in research priorities, it could also be influenced by data incompleteness for ongoing years (especially 2025).

Figure 3 shows the average number of citations per year, which serves as a proxy for the impact and recognition of published work over time.

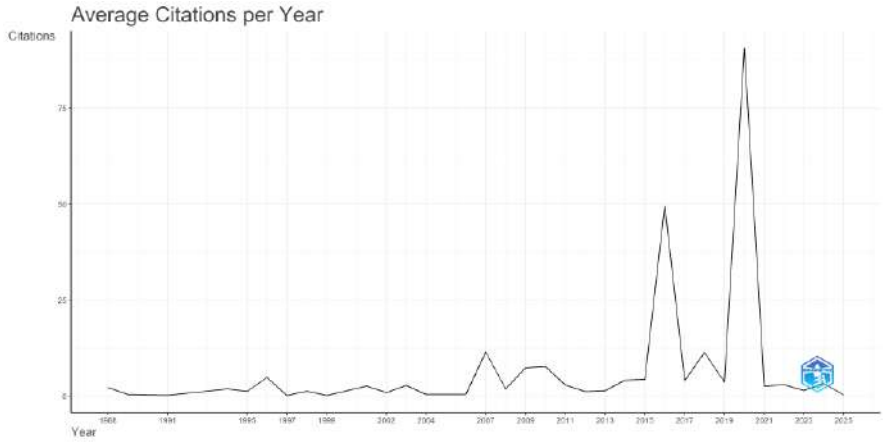

Figure 3. The Average Citation per Year

Citations remained minimal for over two decades. This may be attributed to the niche status of the field or the slow accumulation of scholarly attention. A slight increase is observable, reflecting the gradual integration of earlier works into mainstream literature.

Two notable spikes especially in 2020 suggest the publication of seminal works or highly influential studies that significantly shaped subsequent research. These peaks may correspond to paradigm-shifting articles or widely cited review papers.

A sharp decline in citation averages is visible after 2021. This is a common bibliometric artifact due to the recency of publications: newer articles have had less time to accumulate citations.

2.2 Most Productive Authors and Institutions:

Figure 4 identifies the individual researchers with the highest number of contributions in the dataset:

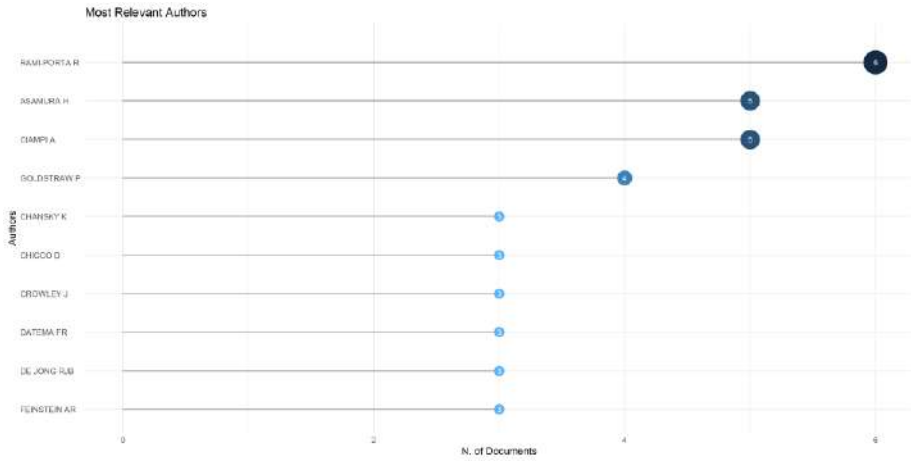

Figure 4. Most Relevant Authors

Rami-Porta, R. leads the author list with six documents, indicating a sustained and influential presence in the literature.

Authors such as Asamura, H., and Ciampi, A. (with 5 publications each), followed by Goldstraw, P. (with 4 publications), suggest a core group of authors actively shaping the discourse within this field. Several scholars, including Chansky, K., Chicco, D., Crowley, J., Datema, F.R., De Jong, R.J.B. and Feinstein, A.R., contributed 3 papers each.

In Figure 5 the visualization of institutional affiliations reflects the distribution of scholarly output by contributing organizations, measured by the number of articles produced:

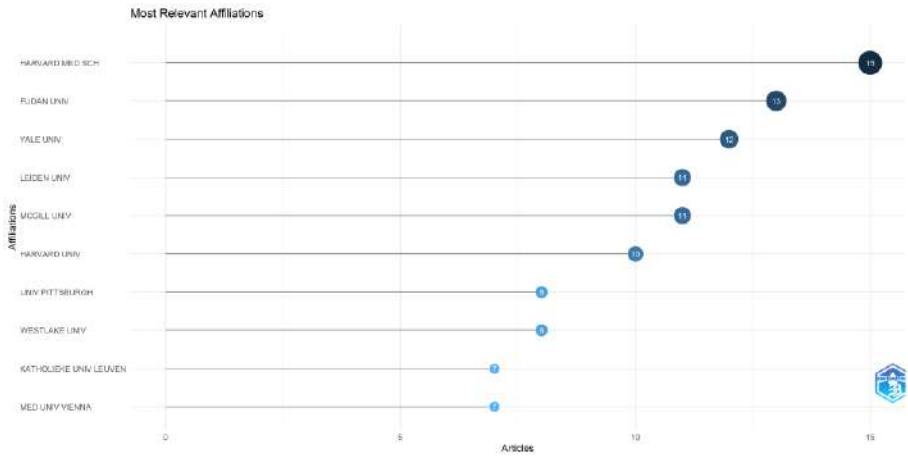

Figure 5. Most Relevant Affiliations

Harvard Medical School stands out with 15 publications, indicating its dominant position and sustained engagement in the field. This suggests a well-established research infrastructure and consistent academic output in the domain. Other high-performing institutions include Fudan University (13 articles), Yale University (12 articles), Leiden University (11 articles), and McGill University (11 articles). The strong presence of universities from North America, Europe, and Asia illustrates a globally distributed research network.

Institutions such as the University of Pittsburgh, Westlake University, Katholieke Universiteit Leuven, and the Medical University of Vienna each contributed between 7 and 8 publications, showing active but comparatively moderate engagement.

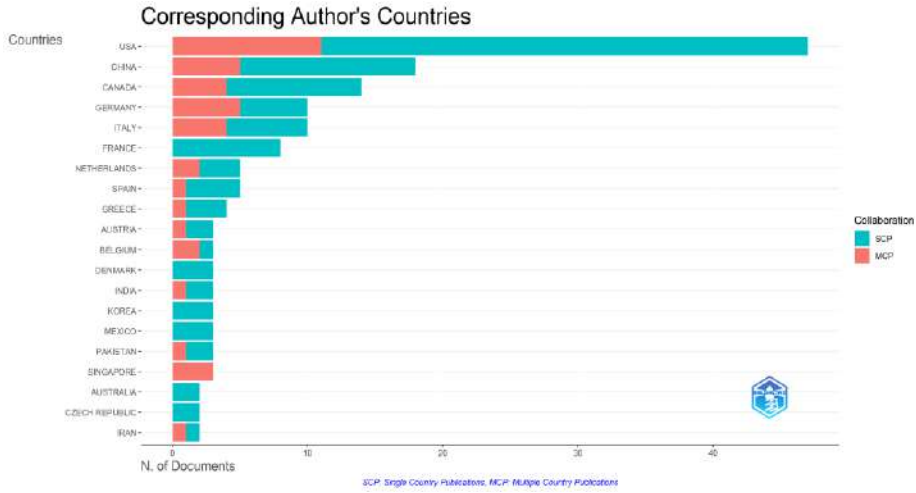

Figure 6. Corresponding Author's Countries

The countries with the highest number of publications are the USA, China, Canada, Germany, and Italy. While most of the publications in the field originated from a single country, all publications in Singapore were published with international cooperation.

2.3 The Most Relevant Sources:

The graph below highlights the distribution of documents across academic journals, providing insight into where the most influential or frequently published works in the field are located.

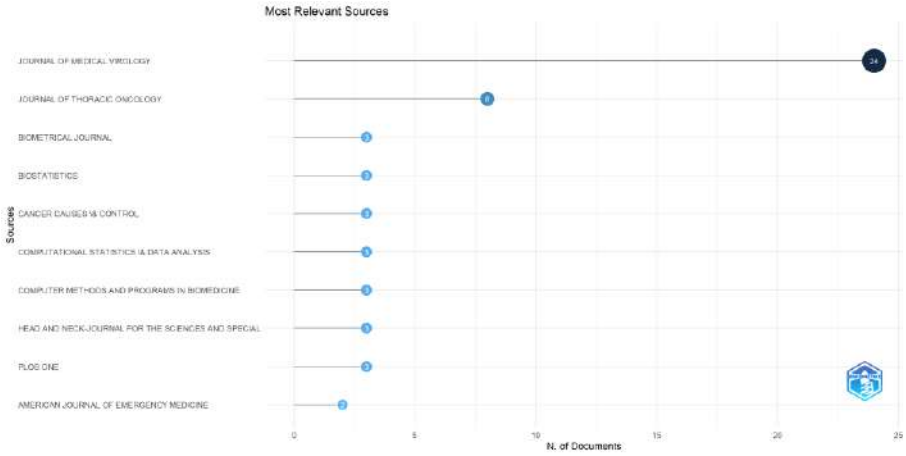

Figure 7. Most Relevant Sources

The Journal of Medical Virology is the most prolific source, with 24 publications, suggesting it serves as a primary outlet for research dissemination in the domain. This dominance may reflect the journal's thematic alignment with the subject area, particularly if it relates to virology, infectious diseases, or epidemiology.

The Journal of Thoracic Oncology follows with eight documents, positioning it as a key specialized journal, likely emphasizing clinical or oncological aspects within the field.

Several other journals including Biometrical Journal, Biostatistics, Cancer Causes & Control, and Computational Statistics & Data Analysis have each published three articles. These venues are strongly associated with methodological rigor and quantitative modeling, indicating a statistically intensive research approach in many contributions.

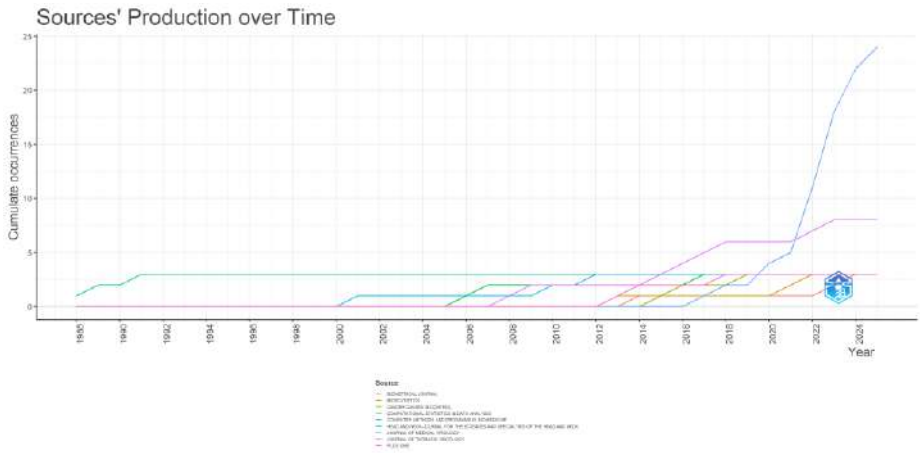

Figure 8. Sources' Production over Time

Figure 8 illustrates the publication output of journals that publish articles on the relevant topic over time. The Journal of Medical Virology, which has the highest number of publications, published its first issue in 1977 and experienced a rapid increase in publications after 2020, driven by the intense demand for virological research during the COVID-19 pandemic. Since 2008, the Journal of Thoracic Oncology has been a leading publication in the field of research, and the number of publications has increased significantly since 2015.

2.4 The Most Cited Articles:

The most highly cited documents in scientific research are the publications that are most cited by other studies in the literature and thus have the highest academic impact. Such documents are not only highly cited but also central in terms of setting the direction of the field,

shaping methodological approaches, and forming the basis for subsequent studies. Therefore, these publications are fundamental contributions to the body of knowledge of the research field. The citations to these studies generally focus on both content and methodological contributions. In this context, the top ten most cited documents in the analyzed study cluster are presented in detail in Table 1.

Table 1: Most Cited Documents

Paper	DOI	TC	TC per Year
Goldstraw P, 2016, J Thorac Oncol	10.1016/j.jtho.2015.09.009	3414	341,40
Chıcco D, 2020, Bmc Genomıcs	10.1186/s12864-019-6413-7	3195	532,50
Sımon R, 2007, Cancer Inform	NA	620	32,63
Park Sh, 2018, Radıology	10.1148/radiol.2017171920	557	69,63
Westreıch D, 2010, J Clın Epıdemıol	10.1016/j.jclinepi.2009.11.020	367	22,94
Eberhardt Wee, 2015, J Thorac Oncol	10.1097/JTO.0000000000000673	311	28,27
Fan J, 2009, Ann Appl Stat	10.1214/08-AOAS215	284	16,71
Ceraolo C, 2020, J Med Vırol	10.1002/jmv.25700	266	44,33
Chıcco D, 2020, Bmc Med Inform Decıs Mak	10.1186/s12911-020-1023-5	243	40,50
Pıccırillo Jf, 1996, Cancer	NA	196	6,53

When the 10 most cited publications are analyzed, the study by Goldstraw et al. 2016) titled "The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer" was the most cited study. Goldstraw and colleagues presented the 8th version of the Tumor-Node-Metastasis (TNM) system for lung cancer and performed Kaplan-Meier survival analyses according to

different TNM combinations, evaluating the prognostic discriminative power of the staging system.

The second most cited study in the field is "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation published by Chicco and Jurman in 2020. In this study, the authors adopt an experimental method to compare the performance measures of classification algorithms. The primary focus is to measure and compare the performance of the Matthews correlation coefficient (MCC) with F1 score and accuracy metrics.

The 3rd most cited paper is "Analysis of Gene Expression Data Using BRB-Array Tools" by Simon et al. in 2007. In this study, gene expression data were classified using Diagonal Linear Discriminant Analysis, Nearest Centroid, Support Vector Machines (SVM), and k-Nearest Neighbors methods.

2.5 Thematic Maps and Keyword Analysis:

The frequency of use of keywords related to classification techniques in biostatistics studies and their changes over time are presented in the figures below. The word "classification" in the literature occupies 13% of the field, indicating that classification problems are at the forefront. Terms such as "prediction," "diagnosis," "survival," "biostatistics," "disease," epidemiology," and "model" are basic methodological and clinical keywords.

Figure 9. Treemap of Keywords

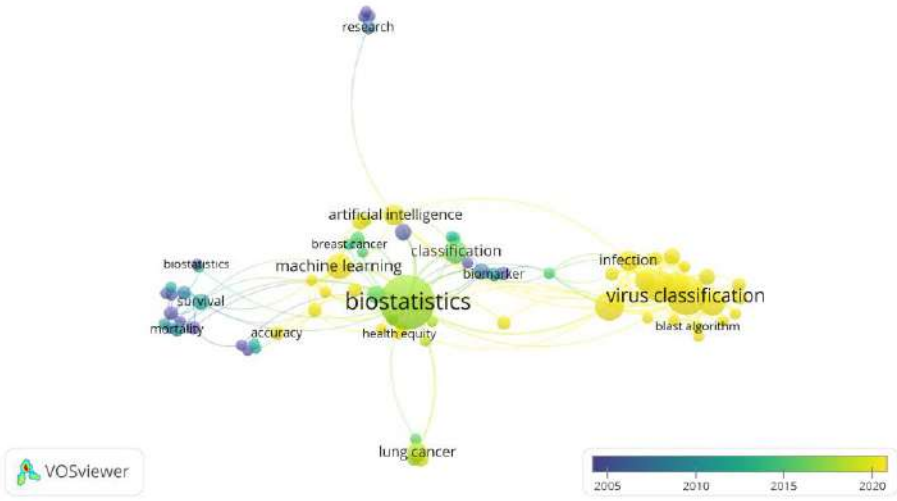

Figure 10. Common asset analysis

The most dominant keyword is "biostatistics," indicating that studies in this field utilize biostatistical methods. Terms such as "survival," "mortality," and "accuracy" were used in early studies from 2010 and before, indicating more classical epidemiological metrics (blue-green tones). Modern analysis techniques such as "machine learning," "artificial intelligence," and "classification" are more recent (yellow tones). The clustering of "virus classification," "infection," and "blast algorithm" indicates a subfield where bioinformatics and infectious diseases are prominent. These groups have become particularly visible in publications since 2020, mainly due to the impact of COVID-19. A shift from classical epidemiological metrics to

artificial intelligence-assisted classification approaches is observed after 2015.

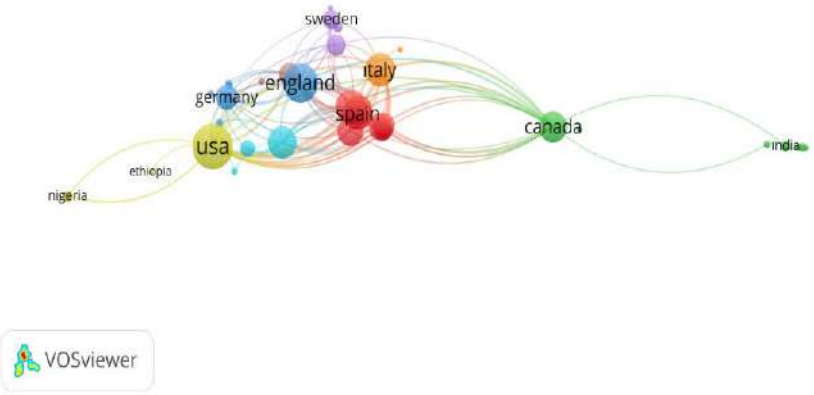

Figure 11. Co-Authorship network by country

According to Figure 10, the country with the most co-authorships is the USA. The USA has established direct collaborations not only with developed countries, such as Germany, the UK, and Canada, but also with developing countries, including Nigeria and Ethiopia. The USA and European countries constitute the center of the relevant literature.

3. DISCUSSION and CONCLUSION

This study employed bibliometric methods to examine the structure, trends, and collaboration networks of scientific production in biostatistics, based on a classification of 170 publications published between 1988 and 2025. The findings shed light on both the historical development of the field and its current scientific dynamics.

Interest in classification methods among academics has increased significantly in recent years. Although a limited number of studies were published prior to 2010, a significant acceleration in annual production was observed after 2015. This increase is primarily attributed to technological advancements in fields such as machine learning, artificial intelligence, bioinformatics, and genomic data analysis. A significant increase in publications on virological classification systems and diagnostic models was observed, particularly during the COVID-19 pandemic.

The analyzed publications had a very high co-authorship rate, prominent international collaborations, and an average author count of over 6.5 per publication. This demonstrates that classification research in biostatistics is essentially a multidisciplinary and collective effort. Countries such as the United States, China, Canada, the United Kingdom, and Germany are at the center of the research network, with developing countries often contributing scientifically through joint projects with these centers.

An examination of the journals in which the publications were published revealed that thematically focused journals, such as the Journal of Medical Virology and the Journal of Thoracic Oncology, stand out, while methodological journals, such as the Biometrical Journal and Biostatistics, offer more limited but impactful contributions. This finding suggests that the field has undergone a two-way development, both in terms of clinical applications and methodological depth.

Keyword analyses also reveal the fundamental conceptual framework of the research. Concepts such as "Classification," "Prediction," "Diagnosis," and "Survival" are prominent, while contemporary terms like "Machine Learning," "Artificial Intelligence," and "Virus Classification" have gained more prevalence since 2020. This trend demonstrates that the field has evolved from classical epidemiological models to modern computational approaches. Finally, an examination of the most cited studies reveals that both publications providing clinical guidelines (e.g., the TNM classification) and those offering methodological evaluations (e.g., a comparison of MCC and F1) have generated high scientific impact. This demonstrates that both content-based and methodological contributions to the field of classification have a lasting impact on the literature.

This bibliometric analysis revealed that classification methods in biostatistics have gained increasing attention over time, and scientific production in this field has accelerated, particularly in the last decade. Classification algorithms have become fundamental tools for

developing diagnostic and predictive models, as well as for systematically evaluating health data.

Research is increasingly utilizing advanced machine learning techniques to analyze the growing volume of data while integrating these approaches with classical statistical methods. This process is increasing interdisciplinary collaboration in the field and contributing to the development of new clinical decision support systems.

In future studies, focusing on comparative success analyses of classification methods, addressing data imbalance problems, and implementing explainable artificial intelligence will improve the quality of both scientific and clinical outputs.

REFERENCES

Archana S, Elangovan K (2014). Survey of Classification Techniques in Data Mining. International Journal of Computer Science and Mobile Applications; 2(2):65-71.

Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics, 11(4), 959–975.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21(1), 6. https://doi.org/10.1186/s12864-019-6413-7.

Fan-Zi, Z., & Zheng-Ding, Q. (2004). A survey of classification learning algorithm. In Proceedings 7th International Conference on Signal Processing, 2004. Proceedings. ICSP'04. 2004. 2:1500-1504. IEEE. https://doi.org/10.1109/ICOSP.2004.1441612.

Fielding, A.H. (2006). Appendix E. In: Cluster and Classification Techniques for the Biosciences. Cambridge University Press; 210-216.

Goel, H., & Kumar, D. (2023, June). Data mining in healthcare using machine learning techniques. In 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS). 25-29. IEEE. https://doi.org/10.1109/ICSCSS57650.2023.10169802.

Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WE, et.al. (2016). The IASLC Lung Cancer Staging

Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. J Thorac Oncol. 11(1):39-51. doi: 10.1016/j.jtho.2015.09.009.

Gupta, M., & Aggarwal, N. (2010). Classification Techniques Analysis. National Conference on Computational Instrumentation.

Jha, S. K., Pan, Z., Elahi, E., & Patel, N. (2019). A comprehensive search for expert classification methods in disease diagnosis and prediction. Expert Systems; 36(1), e12343. https://doi.org/10.1111/exsy.12343.

Kalcheva, N., Todorova, M., & Marinova, G. (2020). Naive bayes classifier, decision tree and adaboost ensemble algorithm–advantages and disadvantages. In Proceedings of the 6th ERAZ Conference Proceedings (part of ERAZ conference collection), 153-157). https://doi.org/10.31410/ERAZ.2020.153.

Kaur, R., Verma, P. (2017). Classification Techniques: A Review. IOSR Journal of Computer Engineering (IOSR-JCE). 19 (1):61-65.

Kesavaraj, G., & Sukumaran, S. (2013, July). A study on classification techniques in data mining. In 2013, the Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (pp. 1-7). IEEE. https://doi.org/10.1109/ICCCNT.2013.6726842.

Khan, H., Srivastav, A. and Mishra, A.K. (2020), "Use of Classification Algorithms in Health Care", Tanwar, P., Jain, V., Liu, C.-M. and Goyal, V. (Ed.) Big Data Analytics and Intelligence: A Perspective for Health Care, Emerald Publishing Limited, Leeds, pp. 31-54. https://doi.org/10.1108/978-1-83909-099-820201007.

Khujaev, O. K., Nurmetova, B. B., & Urazmatov, T. K. (2023). Algorithms for Selecting the Most Efficient Method for Solving Classification Problems. In 2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE). 1740-1743. IEEE. https://doi.org/10.1109/APEIE59731.2023.10347690.

Kourou, K., et al. (2015). Machine Learning Applications in Cancer Prognosis and Prediction Computational and Structural Biotechnology Journal, 13, 8-17. https://doi.org/10.1016/j.csbj.2014.11.005.

Sabouri, Z., Maleh, Y., & Gherabi, N. (2021, November). Benchmarking classification algorithms for measuring the performance on maintainable applications. In The International Conference on Information, Communication & Cybersecurity.173-179. https://doi.org/10.1007/978-3-030-91738-8_17.

Simon, R., Lam, A., Li, M. C., Ngan, M., Menenzes, S., & Zhao, Y. (2007). Analysis of gene expression data using BRB-array tools. Cancer informatics, 3, 117693510700300022.

Thakur, N., & Maheshwari, D.B. (2017). A Review of Image Classification Techniques. Journal of Engineering and Technology (IRJET).

Wang, H., Wassan, J., Zheng, H. (2019). Measurements ıf Accuracy in Biostatistics. Encyclopedia of Bioinformatics and Computational Biology. 1:685-690. https://doi.org/10.1016/B978-0-12-809633-8.20355-5.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523–538.

CHAPTER 3

DETERMINATION OF CERVICAL CANCER BEHAVIORAL RISK FACTORS USING SUPPORT VECTOR MACHINES

Lecturer Dr. Burcu DURMUŞ
Prof. Dr. Öznur İŞÇİ GÜNERİ
Assist. Prof. Dr. Aynur İNCEKIRIK

INTRODUCTION

Cancer, known as the most common disease today, is a growing health problem worldwide. It is crucial for individuals to understand different types of cancer and adopt lifestyle behaviors that help protect them from the disease. Furthermore, early diagnosis is crucial for cancer (İncekırık et al., 2021).

Cervical cancer is one of the most common gynaecological cancers in women and has a high morbidity and mortality rate, especially in developing countries (Sadia, 2022). İncekırık et al. (2021) conducted a study using classification techniques and revealed that gynaecological cancers are much more common in women than in men. According to World Health Organization data, approximately half a million new cases occur each year, and more than 300,000 women die from cervical cancer (Farajimakin, 2024). The most important cause of the disease is high-risk Human Papillomavirus (HPV) infections; however, behavioural and environmental factors such as early sexual initiation, multiple sexual partners, smoking, parity, inadequate

personal hygiene and low socioeconomic status also contribute to increased risk (Kadir et al., 2024; Sadia, 2022).

In today's world, there has been a significant increase in the application of data mining techniques in the healthcare field. Güldoğan et al. (2017) conducted a clinical study demonstrating the performance of support vector machine kernel functions. In their study, conducted for the detection of diabetes, they revealed that support vector machines exhibited high classification performance.

In recent years, the use of machine learning methods in identifying cervical cancer risk factors has increased. These approaches, which overcome the limitations of traditional statistical methods, can more accurately model the relationships between complex risk factors and contribute to early diagnosis/prevention strategies (Ijaz, 2020). In this context, powerful classification algorithms such as Support Vector Machines (SVM) provide results with both high accuracy and generalizability in medical data (Zhang, 2025). The "Cervical Cancer (Risk Factors)" dataset, published by UCI, is widely used for such studies and includes various behavioural, sociodemographic and medical attributes (Dweekat, et al., 2022; UCI, 2017).

The literature demonstrates that SVM-based models exhibit high performance in cervical cancer diagnosis and risk classification. For example, Ijaz (2020) compared SVM with decision trees and extreme learning machine algorithms, reporting that SVM provides higher accuracy in some cases. Similarly, Zhang (2025) demonstrated

that SVM offers better discriminatory performance than other methods in classifying the clinical stages of cervical cancer. Kadir et al. (2024) also reported that an SVM model developed using behavioural risk factors (early sexual intercourse, poor hygiene and poor nutritional habits) was successful in cervical cancer risk prediction.

These studies demonstrate that SVM models developed based on behavioural and social factors can be an important tool in determining cervical cancer risk at an early stage. Therefore, the current study aims to model cervical cancer behavioural risk factors using Support Vector Machines.

MATERIAL AND METHODS

Dataset

The dataset used in the current study is the Cervical Cancer Behavior Risk dataset (Dua & Graff, 2019) from the UCI Machine Learning Repository. This dataset was created to identify behavioural risk factors for cervical cancer and is available to researchers for use in classification studies.

The dataset was donated to UCI on July 16, 2017, and published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license (UCI, 2017). The DOI number is 10.24432/C5402W, and the dataset can be accessed through the UCI Machine Learning Repository.

Dataset structure and general characteristics:

- Total number of observations: 72
- Number of attributes: 20 (19 independent attributes + 1 class label)
- All variables are fully numerical and consist of quantitative data.
- No missing values; data is available for all variables for each observation. There is no missing data.
- Task type: Classification

Variables/Attributes

The attributes in the dataset consist of sub-dimensions of eight main variables, and the first words of their names indicate that main variable. In Table 1 below, each attribute is accompanied by its brief description (based on dataset descriptions).

Table 1. Roles and Descriptions of Variables

	Attribute Name	**Role**	**Description / Note**
1	behavior_sexualRisk	attribute	Risk of sexual behaviour
2	behavior_eating	attribute	Behavioural eating habits
3	behavior_personalHygine	attribute	Personal hygiene behaviour
4	intention_aggregation	attribute	Aggregation intention

5	intention_commitment	attribute	Commitment intention
6	attitude_consistency	attribute	Attitude consistency
7	attitude_spontaneity	attribute	Attitude spontaneity Natural attitude
8	norm_significantPerson	attribute	Significant person's perception of norm
9	norm_fulfillment	attribute	Perception of norm fulfillment
10	perception_vulnerability	attribute	Perceived vulnerability
11	perception_severity	attribute	Perceived disease severity
12	motivation_strength	attribute	Motivation strength
13	motivation_willingness	attribute	Motivation willingness
14	socialSupport_emotionality	attribute	Social support – emotionality
15	socialSupport_appreciation	attribute	Social support – appreciation
16	socialSupport_instrumental	attribute	Social support – instrumental support
17	empowerment_knowledge	attribute	Empowerment – knowledge level

18	empowerment_abilities	attribute	Empowerment – skill/capacity
19	empowerment_desires	attribute	Empowerment – desires/wishes
20	ca_cervix	class	Cervical cancer status (1 = present, 0 = absent)

The class label ca_cervix variable indicates whether the observation is cervical cancer: a value of 1 indicates the presence of cancer and a value of 0 indicates its absence.

Advantages of the Dataset:

- The relatively small size of the data allows for rapid model prototyping.
- The absence of missing values facilitates data cleaning.
- It provides a wide range of risk factors encompassing different behavioural, social and psychological variables.

Limitations of the Dataset:

- Because it contains only 72 observations, generalization capacity may be limited and the risk of overfitting is high.
- The data is drawn from only one sample; therefore, its direct applicability to different populations may be limited.

- Class imbalance is possible (observations with cancer status = 1 may be fewer than those with 0).

Statistical Analysis

Weka 3.9.6 (Waikato Environment for Knowledge Analysis), one of the software programs used for implementing machine learning methods, is widely used in health and biomedical research. This Java-based, open-source software offers numerous functions such as data pre-processing, classification, clustering, attribute selection and model evaluation through its user-friendly interface (Hall, et al., 2009). WEKA, particularly in health data, allows for rapid comparison of different algorithms, making it a prominent tool for epidemiological and clinical data analysis. In the current study, the Support Vector Machines (SVM) algorithm was implemented in the WEKA environment to analyze behavioural risk factors for cervical cancer. Different kernel functions and parameter settings were tested to achieve the highest classification performance. Thus, the integrated analysis environment offered by WEKA software supported the methodological robustness of the study.

Data Pre-processing

Class imbalance is one of the fundamental problems frequently encountered in machine learning applications and classification problems. In this case, when the number of examples belonging to one class is very low compared to the other, the developed model learns the majority class better and tends to ignore the minority class (He & Garcia, 2009). A similar situation exists in the cervical cancer dataset;

the number of observations belonging to the "cancer present" class is significantly lower than the "no cancer present" class. This decreases the sensitivity (recall) of the model, making it difficult to accurately classify individuals with cancer. Therefore, data pre-processing steps to address class imbalance are critical.

To overcome this problem, the SMOTE (Synthetic Minority Over-sampling Technique) method was used in the current study. Instead of directly replicating observations belonging to the minority class, SMOTE generates synthetic examples based on attribute similarities between these observations (Chawla et al., 2002). This approach balances the distribution between classes, allowing the model to better represent the minority class during the training process. It also reduces the risk of overfitting by not adding randomly duplicated examples to the dataset. Thus, Support Vector Machines (SVM) running on a balanced dataset with SMOTE can more reliably predict cervical cancer risk.

Support Vector Machines

Support Vector Machines (SVM) are powerful supervised machine learning algorithms widely used in classification and regression problems. The primary goal of SVM is to find the optimal separating hyperplane (Cortes & Vapnik, 1995). This hyperplane is determined to maximize the margin between classes. The margin is the distance between the data points closest to the hyperplane, called "support vectors". A wider margin generally provides better generalization performance (Cristianini & Shawe-Taylor, 2000).

Support vector machines are one of the most fundamental statistical methods used for classification analysis in data mining. This method is based on predictive logic for linear data and regression logic for nonlinear data (Tezer, 2018). In the support vector machine method, a boundary is drawn to classify the data. This boundary can be drawn in many different ways. The algorithm builds a model based on the line or plane that maximizes classification. Figure 1 shows the support vectors and boundary planes for a two-class problem.

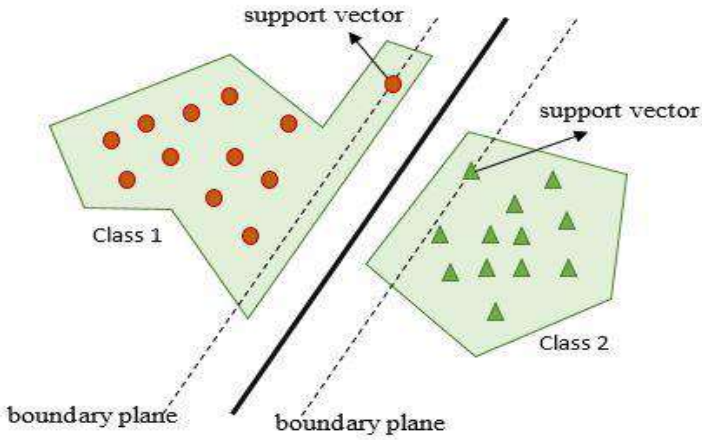

Figure 1. Boundary Planes and Support Vectors (Durmuş & Güneri, 2020).

The primary goal of support vector machines is to approximate the function given by Equation 1.

$$f(x) = \langle w, x \rangle + b$$
$$R_{SVMs}(C) = \frac{1}{2}||w||^2 + C\frac{1}{l}\sum_{i=1}^{l} L_{\varepsilon}(x_i, d_i) \quad (1)$$

Here, $R_{SMSs}(C)$ represents the risk function, $\frac{1}{2}||w||^2$ represents the regularization term and $C\frac{1}{l}\sum_{i=1}^{l} L_{\varepsilon}(x_i, d_i)$ represents the empirical error. The algorithm works with different kernel functions (Durmuş & Güneri, 2020).

SVM is quite effective when the data can be linearly separated. However, in most real-world problems, classes are not linearly separable. In such cases, SVM transforms the data into a higher-dimensional space using kernel functions and attempts to perform linear separation in this space (Schölkopf & Smola, 2002). This approach is called the "kernel trick".

Kernel Functions

Kernel functions are used to calculate the similarity between data and offer different advantages for different problem types. The most commonly used kernel functions in SVM are:

Linear Kernel: This is the simplest kernel function. It is particularly preferred for high-dimensional datasets (e.g., text mining, bioinformatics). Its computational cost is low (Hsu, Chang & Lin, 2010).

Polynomial Kernel: This allows for polynomial modelling of relationships between data. It can represent complex, nonlinear distinctions between classes. The selection of the polynomial degree is critical to the success of the model (Cristianini & Shawe-Taylor, 2000).

Radial Basis Function (RBF Kernel): This is the most widely used kernel function. It creates highly flexible decision boundaries by transforming the data into an infinite-dimensional space. It provides high accuracy, especially for complex problems where classes are not linearly separated (Schölkopf & Smola, 2002).

The most important advantage of SVM is its resistance to overfitting in high-dimensional datasets. Furthermore, it can flexibly adapt to different data distributions by using different kernel functions (Noble, 2006). However, training time can be quite long on large datasets and parameter selections (C, gamma, kernel parameters) significantly affect model performance.

RESULTS

The class distribution in the dataset used in the current study was observed to be unbalanced. Because this imbalance can negatively affect class prediction performance, it was first addressed by applying the SMOTE method. Using SMOTE, the minority class samples were synthetically increased, resulting in a more balanced distribution in the dataset. The results are shown in Figure 2. This aimed to increase the sensitivity of the classification algorithms, particularly on the minority class.

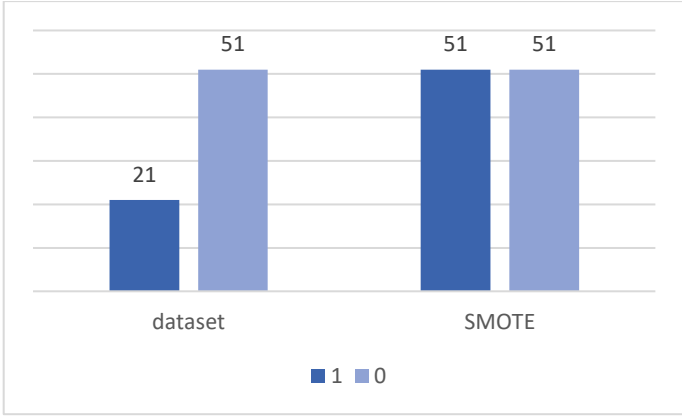

Figure 2. Representation of Data with SMOTE Method Application

In the current study, classification was performed on the class-imbalanced SMOTE dataset using three different Support Vector Machine (SVM) kernels: linear, polynomial and radial (RBF) kernels. The results of the analyses performed with different kernel functions were evaluated in terms of performance criteria such as accuracy, precision, sensitivity and F-measure. The findings and the success levels of the models are presented comparatively in Table 2.

Table 2. SVM Kernel Function Results (Default Selection)

	Linear	**Polynomial**	**Radial (RBF)**
Accuracy	95.098	97.059	86.275
Precision	96.0	94.4	79.4
Recall	94.1	1.0	98.0
F-Measure	95.0	97.1	87.7
MCC	90.2	94.3	74.6
ROC Area	95.1	97.1	86.3
PRC Area	93.3	94.4	78.8
Kappa	0.902	0.941	0.726

Analysis results show that the polynomial kernel (SVM-polynomial) provides the highest performance on the balanced data set. The accuracy (Accuracy = 97.1%), F-Measure (97.1%) and MCC (0.943) values obtained with the polynomial kernel clearly outperformed the other kernels. Furthermore, with Recall = 1.0, all samples in the minority class were successfully classified, demonstrating the advantage of the SMOTE-balanced dataset. The ROC area (97.1%) and Kappa statistic (0.941) values also confirm the superiority of the polynomial kernel in inter-class balance and accurate predictions.

The linear kernel (SVM-linear) demonstrated balanced performance with high precision (96.0%) and accuracy (95.1%). Although the recall value (94.1%) was slightly lower than the polynomial kernel, the overall F-Measure (95.0%) and MCC (0.902) values indicate that the linear kernel is a reliable option. This suggests that the linear kernel may be preferred, especially in scenarios where minimizing false positives is important.

The radial kernel (SVM-rbf) performed lower than the other kernels (Accuracy = 86.3%, F-Measure = 87.7%). Precision (79.4%) and MCC (0.746) values reveal that it is weaker than linear and polynomial kernels in class separation. Although the Recall value (98.0%) was high, some false positive classifications were produced due to the low precision.

In general, datasets balanced with SMOTE significantly improve classification performance and ensure the correct classification

of the minority class. The polynomial kernel stands out as the most suitable kernel for this dataset, providing both high accuracy and balanced class prediction.

DISCUSSION

In the current study, the SMOTE method addressed class imbalance in the dataset containing cervical cancer behavioural risk factors. The SMOTE application balances the inter-class distribution by synthetically increasing the samples in the minority class and improves the model's prediction performance, particularly on the minority class (Chawla et al., 2002; He & Garcia, 2009). Classification was performed on the balanced dataset using three different Support Vector Machine (SVM) kernels - linear, polynomial and radial (RBF) - and performance metrics were analyzed.

The analysis results show that the polynomial kernel (SVM-polynomial) provides the highest performance on the balanced dataset. Accuracy (97.1%), F-Measure (97.1%) and MCC (0.943) values clearly outperformed the other kernels. In particular, the Recall value (1.0) indicates that all samples in the minority class were successfully classified, confirming the effectiveness of the balancing provided by SMOTE. The ROC area (97.1%) and Kappa statistic (0.941) values of the polynomial kernel reinforce its superior performance in inter-class balance and accurate predictions.

The linear kernel (SVM-linear) demonstrated balanced performance with high precision (96.0%) and accuracy (95.1%). While the Recall value (94.1%) was slightly lower than the polynomial kernel, the overall F-Measure (95.0%) and MCC (0.902) values indicate that the linear kernel is a reliable alternative. This suggests that the linear kernel may be preferred, especially in clinical scenarios where minimizing false positives is important.

The radial kernel (SVM-RBF) performed lower than the other kernels (Accuracy = 86.3%, F-Measure = 87.7%). The precision (79.4%) and MCC (0.746) values reveal that it is weaker than the linear and polynomial kernels in class discrimination. Despite the high Recall value (98.0%), the low precision indicates that some false positive classifications occurred.

Analyses reveal that the polynomial kernel performs particularly well in datasets containing high-dimensional and complex relationships. The attributes in the cervical cancer dataset encompass various psychosocial dimensions, such as behavioural (e.g., eating habits, personal hygiene), intention (adherence, collecting), attitude, perception and social support.

The relationships between these attributes are generally non-linear, interactive and complex. The polynomial kernel is capable of modelling these non-linear relationships by transforming the data into a higher-dimensional space (Cristianini & Shawe-Taylor, 2000).

This allows for more successful discrimination between both minority and general class examples.

While the linear kernel captures linear relationships in the data well, it cannot adequately represent the complex interactions between attributes. This resulted in high precision and accuracy, but a slightly lower recall value.

This is reflected in the lower Recall value. From a clinical perspective, the linear kernel's ability to minimize false positives can be useful in screening and preventive strategies, but it may not capture all instances of the minority class (cancer).

The radial basis function (RBF) kernel generally performs well on complex and nonlinear boundaries; however, in this dataset, the low precision and MCC values indicate that the model does not overgeneralize the effects of some attributes.

This suggests that while the radial kernel improves precision for minority class prediction, it is limited in controlling false positives.

Table 3 summarizes the performance metrics, attribute-based contributions and clinical interpretation of each kernel.

Table 3. Results and Interpretations of Kernel Functions

Kernel	Performance Metrics	Attribute-Based Contribution	Clinical Interpretation
Polynomial	Accuracy = 97.1%, F-Measure = 97.1%, MCC = 0.943, Recall = 1.0	Successfully captured complex, nonlinear attribute interactions; behavioural, intention, and social support attributes were fully classified correctly in the model.	The high-impact minority class (cancer present) was correctly classified. The most suitable model for early diagnosis and risk assessment.

Linear	Accuracy = 95.1%, Precision = 96.0%, F-Measure = 95.0%, MCC = 0.902 Recall = 94.1	Successfully captured linear relationships; limited representation of interactions between attributes.	Reliable in clinical scenarios where false positives must be minimized. However, some cancerous samples may be missed.
Radial (RBF)	Accuracy = 86.3%, Precision = 79.4%, F-Measure = 87.7%, MCC = 0.746, Recall = 98.0	It can model high-dimensional data; however, it increased false positives by overgeneralizing some attribute relationships.	Sensitivity for the minority class is high, but false positives are high; additional validation may be required in screening tests.

Consequently, datasets balanced with SMOTE significantly improve classification performance and ensure accurate classification of the minority class. The current study, conducted on the cervical cancer dataset, demonstrates that the polynomial kernel is the most

suitable model, providing both high accuracy and balanced class prediction. This finding highlights the importance of data structure in kernel selection for the clinical use of machine learning-based risk prediction models and can improve the effectiveness of early diagnosis and preventive interventions, especially in health data where behavioural and social attributes are interdependent.

REFERENCES

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. https://doi.org/10.1613/jair.953

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press.

Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Retrieved from https://archive.ics.uci.edu/dataset/537/cervical+cancer+behavior+risk

Durmuş B. ve İşçi Güneri Ö. (2020). Destek Vektör Makinaları ile Erythemato-Skuamöz Hastalıklarının Ayırt Edilmesi ve Çekirdek Fonksiyonlarının Kıyaslanması. 20th Econometrics Operations Research and Statistics Symposium, (pp. 159-165). May 2020. Ankara: Hacı Bayram Veli University. ISBN: 978-605-7893-08-6.

Dweekat, O. Y., Al-Tashi, Q., Rais, H. M., & Alhussian, H. (2022). Cervical cancer diagnosis using an integrated system based on

machine learning algorithms. BioMed Research International, 2022, 1-12.
https://doi.org/10.1155/2022/9601935

Farajimakin, O., Odetunde, O., & Adebola, A. (2024). Barriers to cervical cancer screening: A systematic review. International Journal of Environmental Research and Public Health, 21(3), 456–470. https://doi.org/10.3390/ijerph21030456

Güldoğan, E., Arslan, A.K. ve Yağmur, J., (2017). Çeşitli Çekirdek Fonksiyonları ile Oluşturulan Destek Vektör Makinesi Modellerinin Performanslarının İncelenmesi: Bir Klinik Uygulama, Fırat Medical Journal, 22 (3), 136-142.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter, 11(1), 10–18. https://doi.org/10.1145/1656274.1656278

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hsu, C. W., Chang, C. C., & Lin, C. J. (2010). A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University.

Ijaz, M. F., Attique, M., & Son, Y. (2020). Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. Healthcare, 8(3), 112.

https://doi.org/10.3390/healthcare8030112

İncekırık, A., İşçi Güneri, Ö., & Durmuş, B. (2021). Classification of Cancer Types by Cluster Analysis Methods. Alphanumeric Journal, 9(1), 125-142. https://doi.org/10.17093/alphanumeric.949958

Kadir, K., Rahman, M. M., & Hossain, M. (2024). Predicting cervical cancer based on behavioral risk factors. International Journal of Advanced Computer Science and Applications, 15(11), 1–9. https://thesai.org/Downloads/Volume15No11/Paper_1-Predicting_Cervical_Cancer_Based_on_Behavioral_Risk_Factors.pdf

Noble, W. S. (2006). What is a support vector machine? Nature Biotechnology, 24(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Sadia, H., & Akram, S. (2022). Risk factors of cervical cancer and role of primary prevention. Pakistan Journal of Medical Sciences, 38(3), 554–560. https://doi.org/10.12669/pjms.38.3.4780

Schölkopf, B., & Smola, A. J. (2002). Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT Press.

Tezer, D. (2018) Yapay Sinir Ağları, Lojistik Regresyon ve Destek Vektör Makinesi İstatistik Yöntemlerinin Sınıflandırmadaki Karşılaştırılması, Biruni University, Master's Thesis, Istanbul.

UCI Machine Learning Repository. (2017). Cervical cancer (risk factors). University of California, Irvine. Retrieved from https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+(Risk+Factors)

Zhang, M., Liu, W., & Chen, J. (2025). Prediction of clinical stages of cervical cancer via machine learning models. Frontiers in Oncology, 15, 112345. https://doi.org/10.3389/fonc.2025.112345

ADVANCES IN BIOSTATISTICAL CLASSIFICATION: METHODS, TRENDS, AND MEDICAL APPLICATIONS