Bozok
Tarım ve Doğa Bilimleri Dergisi
(Bozok Journal of Agriculture and Natural Sciences)
https://dergipark.org.tr/en/pub/bojans
e-ISSN: 2822-4604

Araştırma Makalesi

Makine Öğrenmesi Yöntemleri ile Kenevir Türlerinin THC Seviyelerinin Tahmin Edilmesi

Talip ÇAY[1,*]

[1]Yozgat Bozok Üniversitesi, Mühendislik-Mimarlık Fakültesi, Elektrik-Elektronik Mühendisliği Anabilim Dalı, 66100, Yozgat, Türkiye

[1]https://orcid.org/ 0000-0002-5490-1651

*Sorumlu Yazar e-mail: talip.cay@yobu.edu.tr

Makale Tarihçesi

Geliş: 29.10.2024
Kabul: 26.11.2024
DOI: 10.59128/bojans.1575663

Anahtar Kelimeler

Makine öğrenmesi,
k-NN,
SVR,
THC,
Terpen

Öz: Bu çalışmada, hibrit, sativa ve indica kenevir türleri ve terpen profillerine dayanarak THC (tetrahidrokannabinol) seviyelerinin tahmin edilmesi amacıyla makine öğrenmesi algoritmaları kullanılmıştır. Kenevir bitkisi, tıbbi ve endüstri alanında kullanımlarıyla popülarite kazanmıştır; bu nedenle, bu türlerin özelliklerinin anlaşılması kritik öneme sahiptir. Çalışmada, kenevir türlerinin terpen bileşenleri, THC seviyelerini etkileyen önemli faktörlerden biri olarak belirlenmiştir. Veri seti, farklı kenevir türlerine ait terpen bileşenlerini ve bu türlerin THC seviyelerini içermektedir. Çalışmada lineer regresyon, k-NN (k En Yakın Komşuluk), SVR (Destek Vektör Regresyon) ve YSA (Yapay Sinir Ağları) gibi klasik ve derin öğrenme algoritmaları kullanılarak modeller geliştirilmiştir. Her bir algoritma performans metrikleri değerlendirilmiştir. % 95'e ulaşan doğruluk oranları ile elde edilen sonuçlar, terpen bileşenlerinin THC seviyelerinin tahmininde önemli bir rol oynadığını göstermiştir. Bu çalışma, kenevir türlerinin analiz edilmesinde makine öğrenmesi yöntemlerinin uygulanabilirliğini ortaya koymuş ve terpen bileşenleri ile THC seviyesi arasındaki ilişkileri daha iyi anlamak için bir temel oluşturmuştur.

Atıf Künyesi: Çay T. (2024). Makine Öğrenmesi Yöntemleri ile Kenevir Türlerinin THC Seviyelerinin Tahmin Edilmesi, *Bozok Tarım ve Doğa Bilimleri Dergisi,* 3(2), 125-136. **How To Cite:** Çay T. (2024). Prediction of THC Levels in Cannabis Species Using Machine Learning Methods, *Bozok Journal of Agriculture and Natural Sciences,* 3(2), 125-136.

Prediction of THC Levels in Cannabis Species Using Machine Learning Methods

Article Info

Received: 29.10.2024
Accepted: 26.11.2024
DOI: 10.59128/bojans.1575663

Abstract: In this study, machine learning algorithms were used to predict THC (tetrahydrocannabinol) levels based on hybrid, sativa, and indica cannabis lines and their terpene profiles. The cannabis plant has gained popularity due to its medical and industrial applications; therefore, understanding the characteristics of these lines is of critical importance. In the study, the terpene components of cannabis lines were identified as one of the significant factors affecting THC levels. The dataset includes terpene components of different cannabis lines and the THC levels of these lines. Models were developed using classical and deep learning algorithms, such

Keywords

Machine learning,
k-NN,
SVR,
THC,
Terpene

as linear regression, k-NN (k-Nearest Neighbors), SVR (Support Vector Regression), and ANN (Artificial Neural Networks). The performance metrics of each algorithm were evaluated. Results achieved with accuracy rates reaching 95% demonstrated that terpene components play an important role in predicting THC levels. This study highlights the applicability of machine learning methods in analyzing cannabis lines and provides a foundation for better understanding the relationships between terpene components and THC levels.

1. Introduction

Cannabis is one of the oldest agricultural plants in human history, dating back to 2700 BC, and is used for various purposes (Pisanti and Bifulco, 2019). With both industrial and medical uses, cannabis holds significant economic value worldwide (Yildirim and Koca Çalişkan, 2020). The psychoactive effects of cannabis are determined by its various chemicals, particularly cannabinoids such as THC and CBD. THC is a psychoactive compound that induces a feeling of euphoria, while CBD, instead of creating such mental effects, is more noted for its calming and regulatory therapeutic properties (Grotenhermen and Russo, 2013). Additionally, the cannabis plant contains aromatic compounds known as terpenes. Terpenes not only create the characteristic smell of cannabis, but they can also influence potential health benefits (Sommano *et al.*, 2020). Hybrid, sativa, and indica cannabis lines differ in terms of terpene profiles and THC levels (Casano *et al.*, 2011).

Hemp is a fast-growing herbaceous plant that can be grown in different geographical locations and climates. The Sativa lines, which can reach up to 4 meters in height, is notable for its long stem and narrow leaves (Andre, Hausman and Guerriero, 2016). Female plants produce more resin and are therefore richer in chemical components (Campbell, Peach and Wizenberg, 2021). The resin is covered by trichomes found in the flowers of the plant, which are regions rich in psychoactive components such as THC and CBD (Livingston *et al.*, 2020). Sativa lines grow more efficiently in tropical climates due to their wider-spaced branches and longer growth period (Jennings, 1966).

The Indica lines, on the other hand, has a shorter and more compact structure, allowing it to develop quickly even in cold climates and making it suitable for indoor cultivation. Hybrid lines, a cross between Sativa and Indica, display a wide range of physical characteristics, with some hybrids closer to Sativa showing long stems and narrow leaves, while those closer to Indica exhibit short stems and broad leaves. Leaf colors in hybrids can also vary, ranging from light green to dark green and purple tones (Crawford *et al.*, 2021).

THC (Tetrahydrocannabinol) is the main psychoactive component of the cannabis plant, interacting with the cannabinoid receptors in the brain to produce effects on the central nervous system, and its intensity varies depending on the plant's line and cultivation conditions (Costa, 2007). In the medical field, it has been shown to be beneficial in pain management, nausea, and appetite stimulation, and cancer patients can use THC-containing products to alleviate chemotherapy side effects. The terpenes found in cannabis shape not only the plant's aroma and taste but also its therapeutic effects (Hanuš and Hod, 2020). Caryophyllene, known for its peppery aroma, is the only terpene that can directly interact with cannabinoid receptors; limonene improves mood and reduces stress; myrcene, a common terpene in cannabis, has muscle-relaxing and calming properties and is found more in Indica lines; terpinolene has energizing and antioxidant properties; ocimene shows anti-inflammatory and antimicrobial effects; pinene has bronchodilator and anti-inflammatory properties; and linalool, with its lavender-like scent, is known for its calming and sedative effects (Sommano *et al.*, 2020) (Radwan *et al.*, 2021) (LaVigne *et al.*, 2021).

In this context, machine learning and deep learning techniques stand out as powerful tools for analyzing large datasets and making predictions. The aim of the study is to predict THC values for hybrid, sativa, and indica cannabis species based on terpene data and to evaluate the accuracy of these

predictions. Machine learning algorithms such as Regression Analysis, k-NN (k-Nearest Neighbors), SVR (Support Vector Regression), and ANN (Artificial Neural Networks) have been used in the study. The obtained results will be compared, and the effectiveness of these algorithms in cannabis analysis will be examined. Furthermore, the findings of the study will shed light on the relationships between cannabis species and terpene components and will provide suggestions for future research.

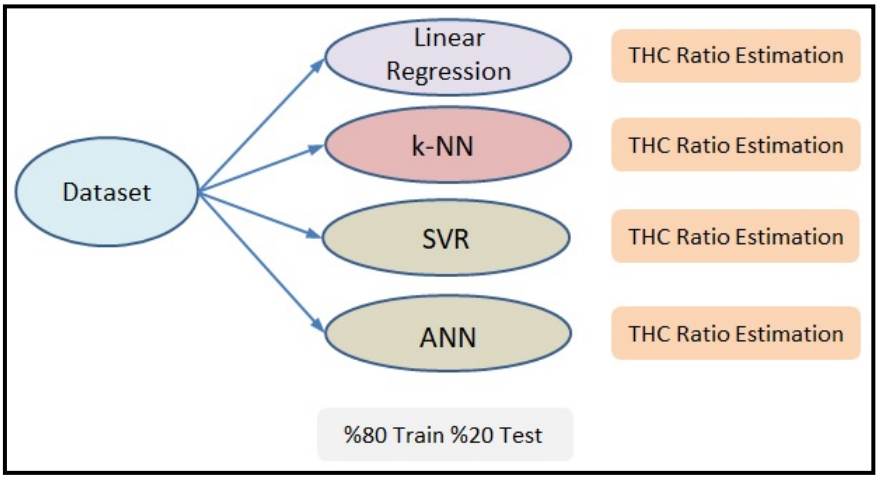

Figure 1. Flow Diagram

1.1. Related Work And Motivation

In the literature, there are numerous studies on the analysis of terpenes and chemical components of the cannabis plant. In our machine learning-based literature review on predicting terpene profiles and THC levels of cannabis plant species, some notable studies are as follows:

ATR-MIR spectroscopy was examined for predicting THC and CBD contents in cannabis flowers and extracts, and it was highlighted as a promising tool for determining product quality standards (Geskovski *et al.*, 2021). The HSI method was developed to predict THCA concentration, with the best results achieved using the RPLS algorithm through sparsity application, marking the first use of HSI in this field (Abeysekera *et al.*, 2023). The relationship between terpenes and cannabinoids in *Cannabis sativa* was analyzed, achieving 99.5% R^2 and 85.71% accuracy with machine learning, demonstrating the predictability of terpenes (Turhan and Yurttakal, 2022). The terpene and cannabinoid profiles of cannabis samples grown indoors and outdoors were compared, revealing significant effects of environmental conditions on chemical components (Zandkarimi *et al.*, 2023). Industrial hemp varieties were classified with 99.6% accuracy using hyperspectral imaging, emphasizing the importance of growth stages (Lu *et al.*, 2022). The inadequacy of current prediction methods was highlighted, and a hybrid methodology was developed to predict cannabis properties, emphasizing the importance of machine learning in chemical engineering (Vergara, Hortúa and Orozco, 2022). The FT-NIRS method was used to rapidly and cost-effectively analyze major cannabinoids and terpenes, achieving over 80% accuracy (Birenboim *et al.*, 2022).

The difference of this study from other studies in the literature lies in the diversity of algorithms used and the high success rate achieved, particularly with Artificial Neural Networks (ANN). In most studies in the literature, the prediction of THC, CBD, and other cannabinoid components is generally aimed using regression analyses and spectroscopic techniques (ATR-MIR, HSI, FT-NIRS). However, these studies often employ simpler regression models (linear regression, PLS) or classification algorithms (LDA, PLS-DA), and success rates vary between 80% and 98%, depending on the techniques used.

In particular, the study compares various algorithms such as linear regression, kNN, SVR, and ANN, achieving a success rate of 95% with ANN. This creates a significant distinction from most studies

in the literature. While the high accuracy rates of ANN generally rely on its ability to learn complex relationships, most studies in the literature have made predictions based on simpler models.

Additionally, the use of ANN enables the more effective modeling of non-linear relationships. This can provide more reliable results in THC prediction based on cannabis species and terpene profiles. While the techniques and algorithms used in many studies in the literature are limited, this study demonstrates a more comprehensive and effective modeling process compared to the approaches in the literature by testing various algorithms and showcasing the high performance of ANN.

As a contribution to future studies, the use of new algorithms, particularly ensemble methods, could enhance the performance of the current study. Ensemble methods allow for stronger and more balanced predictions by combining multiple models. Using larger and more diverse datasets can help the model generalize better under different conditions. Data collected from different regions or climates can improve the model's success in real-world applications. Additionally, transfer learning techniques are an effective way to enhance the model when working with small datasets. Pre-trained models can be adapted to the current study to improve accuracy and reduce training time.

2. Materials and Methods

In this section, detailed information about the machine learning algorithms, performance metrics, and dataset used in the study is presented. The fundamental working principles of the algorithms, along with their advantages and disadvantages, are explained in detail, and how they are applied in prediction processes is elaborated. The performance metrics used in regression analyses to evaluate the accuracy and reliability of the results are also described. Emphasis is placed on how these metrics reflect the quality of the predictions and why they are suitable for this application. Finally, the characteristics of the dataset are presented in detail. This comprehensive explanation aims to ensure the understanding of the methods and data used in the study.

2.1. Machine Learning Methods

In the study, machine learning algorithms such as Linear Regression, k-NN (k-Nearest Neighbors), SVR (Support Vector Regression), and ANN (Artificial Neural Networks) were used.

2.1.1. Linear Regression

Linear regression is a technique used to model the linear relationship between a dependent variable (Y) and one or more independent variables (X). The main goal is to find a linear equation that defines the relationship between Y and X.

Simple linear regression attempts to model the relationship between a dependent variable and a single independent variable. In this case, the regression equation is as follows::

$$Y = a + bx + \epsilon$$

Y, dependent variable (value to be predicted).
X, independent variable (explanatory variable).
a, constant term, which shows the value of Y when X is zero.
b, slope coefficient, which is the effect of a one-unit change in X on the average change in Y.
ϵ, error term is random error or noise that the model cannot explain.

2.1.2. k-NN (k-Nearest Neighborhood)

k-NN (k-Nearest Neighbors) is a simple yet effective algorithm used in machine learning to classify observations or make value predictions. The basic principle of this method is to predict by comparing a new observation point with its k nearest neighbors in the dataset. For classification, the class of the observation is determined based on the majority of class labels of the k nearest neighbors.

In regression analysis, prediction is made by taking the average of the target variable values of the k nearest neighbors (Hastie, Tibshirani and Friedman, 2009) (James *et al.*, 2023).

The first step in classifying or predicting a new observation point is to calculate the distances between this observation and all observations in the training dataset. The most commonly used method for measuring these distances is Euclidean distance, which is preferred to calculate the straight-line distance between two points.

$$d(x_i, x_j) = \sqrt{\sum_{p=1}^{n} (x_{ip} - x_{jp})^2}$$

k-NN distance calculation

Here xi and xj represent the new observation and an observation in the training dataset, respectively.

After the neighbors are determined, the distances between the new observation and the other points in the training dataset are ranked to find the k nearest neighbors.

The main advantages of the k-NN algorithm include its simplicity, flexibility, and ease of application. However, it also has disadvantages, such as the potential increase in computational cost when working with large datasets and the risk that irrelevant features may negatively affect the model's performance. In the study, the optimal k value was chosen as 5. Euclidean distance was used as the distance measurement.

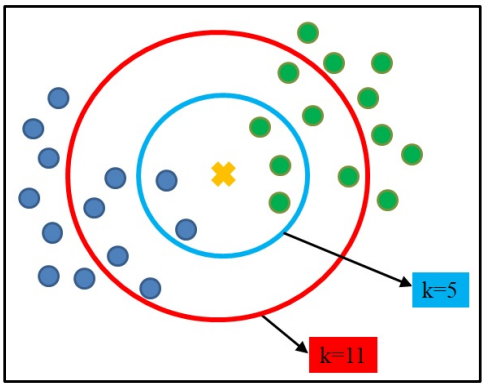

Figure 2. Determining the k value

2.1.3. SVR (Support Vector Regression)

SVR (Support Vector Regression) is a machine learning algorithm commonly used to solve regression problems. It tries to optimize both fitting to the training data and its generalization ability. SVR uses a loss function called epsilon-insensitive loss function to make predictions for the target variable.

SVR essentially follows these steps: The first step involves separating the data points with a hyperplane, ensuring that the data remains within a certain margin of error for accurate predictions. The second main step places the data points within a margin, allowing SVR to expose errors within the given tolerance (epsilon) range (Awad and Khanna, 2015).

$$f(x) = w.\emptyset(x) + b$$

SVR target function

Where; w is a linear vector, b is the bias term and φ(x) is the kernel function that transforms the data points into a high-dimensional space.

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad \text{subject to } |y_i - f(x_i)| \leq \epsilon + \varepsilon_i, \forall_i$$

SVR loss function

Here; y_i is the target value, $f(x_i)$ is the predicted value of the model, ε_i is the error term.

SVR is effectively used in time series forecasting, price predictions in financial markets, weather forecasting, and other analyses. It is also widely preferred in applications such as modeling chemical processes and predicting molecular properties in the chemical and biotechnology industries. In engineering and industrial fields, successful results can be achieved in production processes, quality control, and robotic systems. Due to this versatility, SVR is a valuable tool in many sectors (Vilela *et al.*, 2019).

In the study, RBF was selected as the kernel function for SVR parameters, C value was taken as 1 and gamma as 0.1.

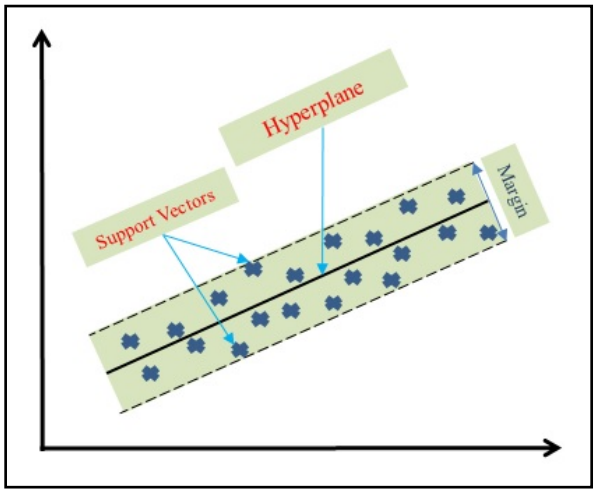

Figure 3. Support Vector Regression (SVR)

2.1.4. ANN (Artificial Neural Networks)

ANN (Artificial Neural Networks) is a subfield of artificial intelligence that enables algorithms to match learned models with data. While it works similarly to traditional machine learning algorithms, deep ANN deals with more complex structures and larger datasets. ANN has a multi-layered architecture, where data passes through different layers to learn features, and in the final layer, classification, prediction, or another output is produced. These features allow the model to understand the data and distinguish between different elements.

In the 1950s, inspired by the works of Frank Rosenblatt, Warren McCulloch, and Walter Pitts, the simplest ANN model, the Perceptron, was developed. Although the Perceptron was successful in simple tasks, it proved inadequate in solving complex problems. In the early 1990s, advancements in machine learning led to further development of ANN, and new learning techniques, such as the backpropagation algorithm, were introduced. This enabled ANN to perform more complex tasks. In the early 2000s, there was a significant increase in the use of deep ANN with multiple hidden layers. Today, deep ANN is considered one of the most important tools in machine learning and is used in various areas of healthcare, such as disease diagnosis and medical image interpretation (Lu *et al.*, 2024).

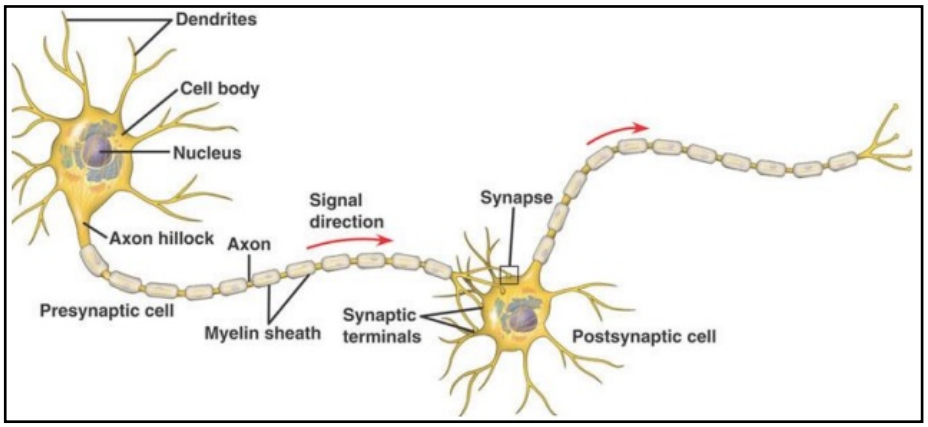

Figure 4. Biological Neuron (Hitziger and Modeling, 2015)

As seen in Figure 4, there is a nucleus in the nerve cell. Data emerging from the nucleus travels along the axon and is transmitted to other nerve cells through the axon terminals, which are connected to the dendrites of another nerve cell. The points where the axon terminals and dendrites are connected are called synapses. The data transmitted from other nerve cells to the nucleus is processed in the nucleus and then transmitted to other nerve cells.

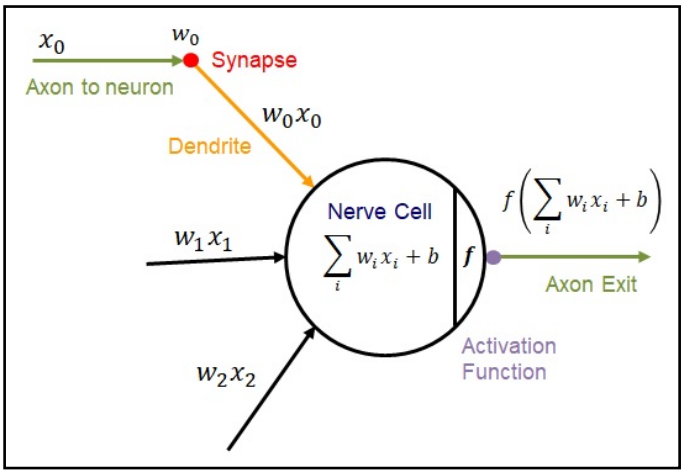

Figure 5. Mathematical Modeling of an Artificial Neuron

The mathematical modeling of the data processing and transmission process of a biological nerve cell is shown in Figure 5. Here, the input information from another artificial nerve cell is multiplied by weights and transmitted to the nerve cell. The counterparts of these weights in biological nerve cells are the dendrites. The input information multiplied by the weights is summed, and a bias value is added. Then, the total value is passed through an activation function, and the data is transmitted to the artificial nerve cell in the next layer. This is how the artificial nerve cell operates.

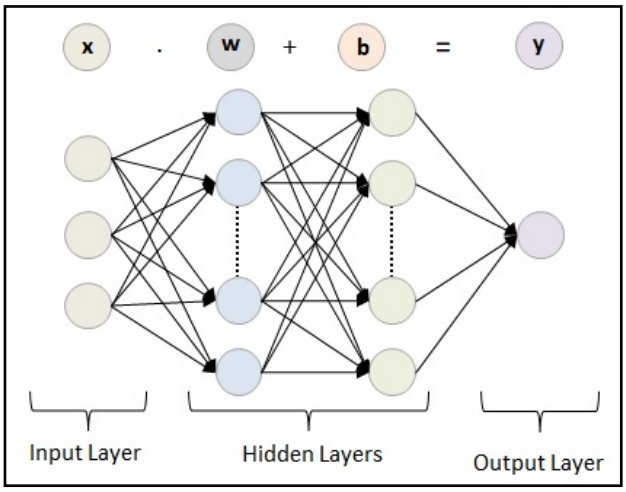

Figure 6. ANN Model

The data value of each nerve cell is calculated, and as shown in Figure 6, nerve cells are connected in both serial and parallel configurations. The neural network consists of input, output, and hidden layers. Neural networks with multiple hidden layers are referred to as multilayer neural networks (deep neural networks). If the neural network has a single hidden layer, it is referred to as a single-layer neural network.

Sequence Input Layer
Fully Connected Layer (32)
ReLU Layer
Fully Connected Layer (64)
ReLU Layer
Fully Connected Layer (64)
ReLU Layer
Fully Connected Layer (128)
ReLU Layer
Dropout Layer
Fully Connected Layer (128)
ReLU Layer
Dropout Layer
Fully Connected Layer (2)
Softmax Layer
Classification Output Layer

Figure 7. Designed network layers

The deep neural network model used in this study consists of 16 layers, as shown in Figure 7. The ReLU activation function has been chosen for the model. The model training was performed using the ADAM optimization algorithm for 1000 epochs.

2.2. Performance Metrics

The following metrics are generally used to evaluate the performance of regression methods;

MAE (Mean Absolute Error) is the average of the absolute differences between the predicted values and the actual values. A low MAE indicates that the model's predictions are close to the actual values.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean absolute error
n, number of samples in the data set
y_i, real values
$\hat{y}_i$, estimated values

MSE (Mean Squared Error) is the average of the squared differences between the actual values and the predicted values.

$$\boldsymbol{MSE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\mathbf{2}}$$

Mean squared error
RMSE (Root Mean Squared Error) is the square root of MSE. Similar to MSE but ensures that the units are the same. It has more impact on large errors.

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

R-squared (R^2) is commonly used to measure the success of a model. It represents the relationship between the actual values and the model's predictions. The closer it is to 1, the more accurate the model's predictions are.

$$\boldsymbol{R^2} = \mathbf{1} - \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i}-\boldsymbol{\hat{y}_i})^{\mathbf{2}}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i}-\boldsymbol{\bar{y}})^{\mathbf{2}}}$$

$\bar{y}$, is the average of the true values.

2.3. Dataset Properties

The study uses data from hybrid, sativa, and indica cannabis lines. The dataset consists of 2173 samples. The dataset includes key variables such as cannabis line, terpene profiles, and THC levels. As shown in Figure 9, the terpenes include various components such as Caryophyllene, Limonene, Myrcene, Terpinolene, Ocimene, Pinene, Linalool, and Humulene. The dataset comprehensively documents the characteristics specific to each cannabis line. Notably, key terpenes such as Caryophyllene, Limonene, Myrcene, and Terpinolene stand out as the main components between the lines. These terpenes are highlighted as prominent compounds in each cannabis line and play a critical role in understanding the biochemical profile of cannabis and the potential effects of the lines. The THC levels of the cannabis lines in the dataset range from 13% to 28%. Additionally, the dataset is split into 80% training and 20% testing data for the training and testing procedures.

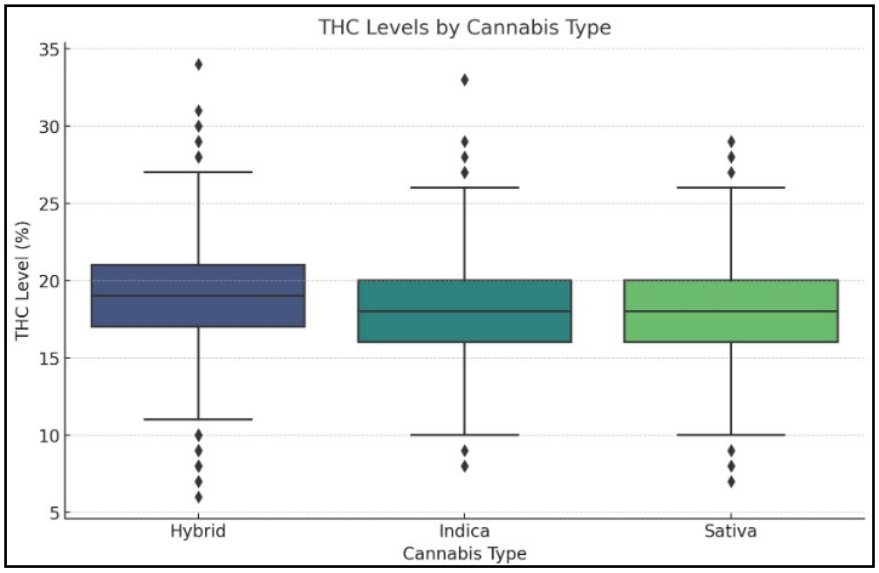

Figure 8. Distribution of THC levels by species

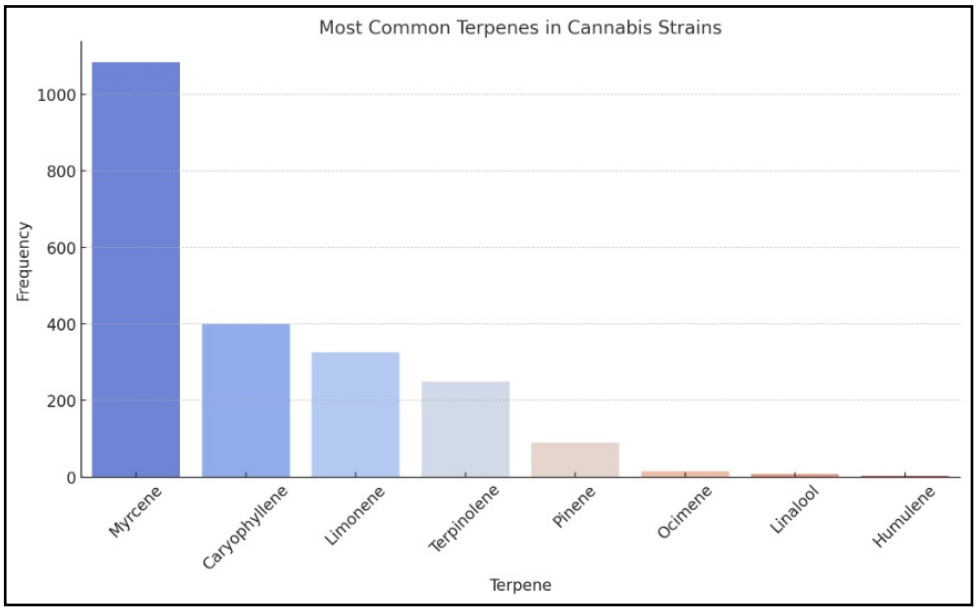

Figure 9. Distribution of terpenes

3. Conclusion and Evaluation

This study compared the performance of classical and deep learning algorithms in predicting THC values using terpene profiles of hybrid, sativa, and indica cannabis lines. The K-Nearest Neighbors (KNN) algorithm showed a relatively successful performance with an accuracy rate of 89%, but its limitations may arise as the data becomes more complex. The Support Vector Regression (SVR) algorithm achieved a higher accuracy rate of 90%, demonstrating its capacity for more precise differentiation on high-dimensional data. Linear regression showed a high performance with an accuracy rate of 92%, demonstrating that cannabis lines can be predicted based on their terpene compositions. However, the accuracy rate may be limited in datasets containing more complex or non-linear relationships, necessitating the use of more advanced algorithms. Artificial neural networks (ANN) achieved the highest success with an accuracy rate of 95%, demonstrating its ability to learn more complex and qualitative relationships. The results suggest that terpene components are a significant factor in predicting THC levels and contribute to understanding their effects on the biological and genetic diversity of cannabis. Future research should investigate different cannabis lines across broader and more diverse datasets and conduct more in-depth analyses of the effects of terpene components on THC and other cannabinoid levels.

Table 1. Results

Model	Accuracy (%)	R^2	MAE	MSE	RMSE
Linear Regression	92	0.9246	0.1450	0.0236	0.1535
kNN	89	0.8985	0.1675	0.0317	0.1780
SVR	90	0.9020	0.1625	0.0306	0.1750
ANN	95	0.9574	0.1125	0.0133	0.1154

4. Data Availability Statement

The data used in our study are publicly available and not subject to data privacy, as detailed in our Data Availability Statement. Therefore, there is no requirement for ethical approval for the use of

these data. (Cannabis Lines Dataset. Kaggle. https://www.kaggle.com/datasets/corykjar/leafly-cannabis-lines-dataset [Access date: October 10, 2024])

5. References

Abeysekera, S.K. *et al.* (2023) 'Sparse reproducible machine learning for near infrared hyperspectral imaging: Estimating the tetrahydrocannabinolic acid concentration in Cannabis sativa L.', *Industrial Crops and Products*, 192, p. 116137. Available at: https://doi.org/10.1016/J.INDCROP.2022.116137.

Andre, C.M., Hausman, J.F. and Guerriero, G. (2016) 'Cannabis sativa: The plant of the thousand and one molecules', *Frontiers in Plant Science*, 7(FEB2016), p. 174167. Available at: https://doi.org/10.3389/FPLS.2016.00019/BIBTEX.

Awad, M. and Khanna, R. (2015) 'Support Vector Regression', *Efficient Learning Machines*, pp. 67–80. Available at: https://doi.org/10.1007/978-1-4302-5990-9_4.

Birenboim, M. *et al.* (2022) 'Use of near-infrared spectroscopy for the classification of medicinal cannabis cultivars and the prediction of their cannabinoid and terpene contents', *Phytochemistry*, 204, p. 113445. Available at: https://doi.org/10.1016/J.PHYTOCHEM.2022.113445.

Campbell, L.G., Peach, K. and Wizenberg, S.B. (2021) 'Dioecious hemp (Cannabis sativa L.) plants do not express significant sexually dimorphic morphology in the seedling stage', *Scientific Reports 2021 11:1*, 11(1), pp. 1–8. Available at: https://doi.org/10.1038/s41598-021-96311-w.

Casano, S. *et al.* (2011) 'Variations in terpene profiles of different strains of Cannabis sativa L.', *Acta Horticulturae*, 925, pp. 115–122. Available at: https://doi.org/10.17660/ACTAHORTIC.2011.925.15.

Costa, B. (2007) 'On the Pharmacological Properties of Δ9-Tetrahydrocannabinol (THC)', *Chemistry & Biodiversity*, 4(8), pp. 1664–1677. Available at: https://doi.org/10.1002/CBDV.200790146.

Crawford, S. *et al.* (2021) 'Characteristics of the Diploid, Triploid, and Tetraploid Versions of a Cannabigerol-Dominant F1 Hybrid Industrial Hemp Cultivar, Cannabis sativa "Stem Cell CBG"', *Genes 2021, Vol. 12, Page 923*, 12(6), p. 923. Available at: https://doi.org/10.3390/GENES12060923.

Geskovski, N. *et al.* (2021) 'Mid-infrared spectroscopy as process analytical technology tool for estimation of THC and CBD content in Cannabis flowers and extracts', *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 251, p. 119422. Available at: https://doi.org/10.1016/J.SAA.2020.119422.

Grotenhermen, F. and Russo, E. (2013) 'Cannabis and Cannabinoids: Pharmacology, Toxicology, and Therapeutic Potential', *Cannabis and Cannabinoids: Pharmacology, Toxicology, and Therapeutic Potential*, pp. 1–439. Available at: https://doi.org/10.4324/9780203479506/CANNABIS-CANNABINOIDS-ETHAN-RUSSO-ETHAN-RUSSO.

Hanuš, L.O. and Hod, Y. (2020) 'Terpenes/Terpenoids in Cannabis: Are They Important?', *Medical Cannabis and Cannabinoids*, 3(1), pp. 25–60. Available at: https://doi.org/10.1159/000509733.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) 'The Elements of Statistical Learning'. Available at: https://doi.org/10.1007/978-0-387-84858-7.

Hitziger, S. and Modeling, S.H. (2015) 'Modeling the variability of electrical activity in the brain'. Available at: https://theses.hal.science/tel-01175851 (Accessed: 29 October 2024).

James, G. *et al.* (2023) 'An Introduction to Statistical Learning'. Available at: https://doi.org/10.1007/978-3-031-38747-0.

Jennings, P.R. (1966) *The Evolution of Plant Type in Oryza sativa on JSTOR*, *Economic Botany*. Available at: https://www.jstor.org/stable/4252799 (Accessed: 29 October 2024).

LaVigne, J.E. *et al.* (2021) 'Cannabis sativa terpenes are cannabimimetic and selectively enhance cannabinoid activity', *Scientific Reports 2021 11:1*, 11(1), pp. 1–15. Available at: https://doi.org/10.1038/s41598-021-87740-8.

Livingston, S.J. *et al.* (2020) 'Cannabis glandular trichomes alter morphology and metabolite content during flower maturation', *The Plant Journal*, 101(1), pp. 37–56. Available at: https://doi.org/10.1111/TPJ.14516.

Lu, S. *et al.* (2024) 'Advances in Machine Learning Processing of Big Data from Disease Diagnosis Sensors', *ACS Sensors*, 9(3), pp. 1134–1148. Available at: https://doi.org/10.1021/ACSSENSORS.3C02670/ASSET/IMAGES/MEDIUM/SE3C02670_0010.GIF.

Lu, Y. *et al.* (2022) 'Hyperspectral Imaging With Machine Learning to Differentiate Cultivars, Growth Stages, Flowers, and Leaves of Industrial Hemp (Cannabis sativa L.)', *Frontiers in Plant Science*, 12, p. 810113. Available at: https://doi.org/10.3389/FPLS.2021.810113/BIBTEX.

Pisanti, S. and Bifulco, M. (2019) 'Medical Cannabis: A plurimillennial history of an evergreen', *Journal of Cellular Physiology*, 234(6), pp. 8342–8351. Available at: https://doi.org/10.1002/JCP.27725.

Radwan, M.M. *et al.* (2021) 'Cannabinoids, Phenolics, Terpenes and Alkaloids of Cannabis', *Molecules 2021, Vol. 26, Page 2774*, 26(9), p. 2774. Available at: https://doi.org/10.3390/MOLECULES26092774.

Sommano, S.R. *et al.* (2020) 'The Cannabis Terpenes', *Molecules 2020, Vol. 25, Page 5792*, 25(24), p. 5792. Available at: https://doi.org/10.3390/MOLECULES25245792.

Turhan, N. and Yurttakal, A.H. (2022) 'Regression and Discrimination Performance of Terpenoid Expression via Cannabis Sativa Cannabinoids', *HORA 2022 - 4th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings* [Preprint]. Available at: https://doi.org/10.1109/HORA55278.2022.9800044.

Vergara, L.A., Hortúa, H.J. and Orozco, G.A. (2022) 'Property Estimation Method for Cannabinoids and Terpenes Using Machine Learning', *Computer Aided Chemical Engineering*, 51, pp. 103–108. Available at: https://doi.org/10.1016/B978-0-323-95879-0.50018-7.

Vilela, L.F.S. *et al.* (2019) 'Forecasting financial series using clustering methods and support vector regression', *Artificial Intelligence Review*, 52(2), pp. 743–773. Available at: https://doi.org/10.1007/S10462-018-9663-X/FIGURES/11.

Yildirim, S. and Koca Çalişkan (2020) 'Hemp And Its Use In Health (Kenevir Ve Sağlık Alanında Kullanımı)', *avesis.gazi.edu.tr U Koca Çalışkan J. Fac. Pharm. Ankara/Ankara Ecz. Fak. Derg., 2020•avesis.gazi.edu.tr*, 44(1), pp. 112–136. Available at: https://doi.org/10.33483/jfpau.559665.

Zandkarimi, F. *et al.* (2023) 'Comparison of the Cannabinoid and Terpene Profiles in Commercial Cannabis from Natural and Artificial Cultivation', *Molecules*, 28(2), p. 833. Available at: https://doi.org/10.3390/MOLECULES28020833/S1.