A Machine Learning Approach to Evaluating the Impact of Natural Oils on Alzheimer’s Disease Progression

Rema M. Amawi 1,*, Khalil Al-Hussaeni 2,*, Joyce James Keeriath 3 and Naglaa S. Ashmawy 4,5,*

1 Mathematics and Sciences Department, Rochester Institute of Technology, Dubai 341055, United Arab Emirates
2 Computing Sciences Department, Rochester Institute of Technology, Dubai 341055, United Arab Emirates
3 Electrical Engineering Department, Rochester Institute of Technology, Dubai 341055, United Arab Emirates; jjk9192@rit.edu
4 Pharmaceutical Sciences Department, College of Pharmacy, Gulf Medical University, Ajman 4184, United Arab Emirates
5 Pharmacognosy Department, Faculty of Pharmacy, Ain Shams University, Cairo 11566, Egypt
* Correspondence: rxacada@rit.edu (R.M.A.); kxacad@rit.edu (K.A.-H.); dr.naglaa@gmu.ac.ae or naglaa.saad@pharma.asu.edu.eg (N.S.A.)

Abstract: Alzheimer’s Disease is among the major chronic neurodegenerative diseases that affects more than 50 million people worldwide. This disease irreversibly destroys memory, cognition, and the overall daily activities which occur mainly among the elderly. Few drugs are approved for Alzheimer’s Disease management despite its high prevalence. To date, the available drugs in the market cannot reverse the damage of neurons caused by the disease leading to the exacerbation of symptoms and possibly death. Medicinal plants are considered a rich source of chemical constituents and have been contributing to modern drug discovery in many therapeutic areas including cancer, infectious, cardiovascular, neurodegenerative and Central Nervous System (CNS) diseases. Moreover, essential oils that are extracted from plant organs have been reported for a wide array of biological activities, and their roles as antioxidants, antiaging, cytotoxic, anti-inflammatory, antimicrobial, and enzyme inhibitory activities. This article highlights the promising potential of plants’ essential oils in the discovery of novel therapeutic options for Alzheimer’s Disease and halting its progression. In this article, 428 compounds were reported from the essential oils isolated from 21 plants. A comparative study is carried out by employing a variety of machine learning techniques, validation, and evaluation metrics, to predict essential oils’ efficacy against Alzheimer’s Disease progression. Extensive experiments on essential oil data suggest that a prediction accuracy of up to 82% can be achieved given the proper data preprocessing, feature selection, and model configuration steps. This study underscores the potential of integrating machine learning with natural product research to prioritize and expedite the identification of bioactive essential oils that could lead to effective therapeutic interventions for Alzheimer’s Disease. Further exploration and optimization of machine learning techniques could provide a robust platform for drug discovery and development, facilitating faster and more efficient screening of potential treatments.

Keywords: Alzheimer’s disease treatments; essential oils; natural oils; data mining; prediction

1. Introduction

Traditionally, natural products have been used for the treatment of many diseases [1]. Medicinal plants are among the major sources of traditional medicines. Moreover, several modern medicines are produced indirectly from medicinal plants [2]. Natural products have played an important role in drug discovery in many therapeutic areas, including cancer [3], infectious diseases [4], cardiovascular [5] and CNS diseases [6]. Due to the unique chemical diversity of natural products, their biological activities and drug-like properties are very diverse in comparison to synthetic drugs. Natural products...
have served as a structural pool for small molecular libraries to discover biologically active drugs [6]. Nowadays, natural products including medicinal plants have been considered as main targets in drug-discovery programs [7].

Plant essential oils are well known in traditional medicine as well as aromatherapy for the management of numerous diseases [8]. Essential oils are mixtures of various volatile chemical compounds and are extracted from different plant organs, such as roots, leaves, and fruits. Their biological activities have been studied and reported in the literature for decades and their popularity has only increased over time. Examples of such activities include antioxidants, antiaging, cytotoxic, anti-inflammatory, antimicrobial, and enzyme inhibitory activities [9–11]. Essential oils have also been reported as promising agents for the treatment of neurodegenerative diseases due to possessing strong free radical scavenging activity, and hence inhibition of oxidative stress in the body [12].

Alzheimer’s Disease is considered to be the number one chronic neurodegenerative disease, that affects more than 50 million people worldwide [13]. Being a neurodegenerative disease, it slowly and irreversibly destroys memory, cognition, as well as the ability to perform daily activities, resulting in the requirement for full-time care, which occurs mainly among individuals over 65 [14]. Being the most common cause of dementia, Alzheimer’s Disease is considered the third major death cause for the elderly, after cardiovascular diseases and cancer [15].

Despite the high prevalence of Alzheimer’s Disease, only five drugs have been approved by the Food and Drug Administration (FDA) for its management, namely galantamine, rivastigmine, donepezil, memantine, and the combination of memantine and donepezil. Moreover, none of the available drugs can reverse or stop the damage of neurons which causes Alzheimer’s Disease symptoms leading to disease mortality [16].

Among the treatment approaches for Alzheimer’s Disease is the use of choline esterase inhibitors, one of the FDA-approved drugs galantamine which inhibits the choline esterase enzyme is a natural product [17]. In this work, we biologically screen the targeted and isolated pure compounds, in addition to the selected essential oils, in order to identify potential active compounds against Alzheimer’s Disease progression.

Some essential oils have been documented to improve memory and learning abilities. Using essential oils as a potential source of treatment for Alzheimer’s could achieve improvement of patients’ quality of life in terms of cognitive abilities, mental health, and social interactions [18,19]. We study the components of 21 natural oils commercially available in the United Arab Emirates and examine their effectiveness in preventing Alzheimer’s Disease progression. Based on the amount of research interest in these oils; Amla, Anjeer, Apricot, Avocado, Chamomile, Costus, Ginger, Ginseng, Grapefruit, Gum Myrrh, Hazelnut, Henna, Juniper, Mint, Mustard, Onion, Rosemary, Sadab, Sandal, Sweet Violet, Turmeric, they were chosen to be the focus of this study to enable comparison with previous results.

Machine learning (ML) has been increasingly prevalent in the scientific literature due to its ability to analyze vast amounts of data and produce accurate results. Using ML techniques, we develop a model trained on the composition of the 21 oils and their activity against Alzheimer’s progression. The purpose of this model is to predict whether a new and previously untested oil could have potential activity against Alzheimer’s as well, which may save time in terms of laboratory testing. The potential treatments could then be provided to Alzheimer’s patients, thus offering a cost-effective solution for this age-old disease.

The paper is organized as follows. Section 2 surveys the literature for related work. Section 3 details our proposed methodology, including the employed dataset, data preparation and preprocessing, and building machine learning models. The experimental settings and results are detailed in Section 4. A discussion of the results and the application of our work is presented in Section 5. Finally, the paper concludes in Section 6.
2. Literature Review

We conduct a thorough literature review on relevant natural oils to this study. We survey the body of existing research work on the impact of natural oils in the medical domain while shedding light on effective machine-learning techniques used in such studies.

An interesting study by Abdel-Hady et al. in 2022 highlighted the benefits of Amla in the treatment of nausea, asthma, bronchitis, leucorrhoea and vomiting, in addition to having antipyretic and anti-inflammatory properties [20]. Anjeer, commonly known as Fig, was studied by a group of researchers from India in 2014 who reported antioxidant and antibacterial activities and highlighted the potential of nutritional and therapeutic benefits [21,22].

Nafis et al. in 2020 stated that apricots exhibited antimicrobial activity and can be considered for their potential in combating multidrug-resistant strains [23], as well as Costus which displayed a similar activity according to Shafti et al. in 2015 [24]. Avocado is one of the most researched fruits as evident in the hundreds of articles published over the last two decades. In addition to its popular use in cosmetics and culinary industries, it has shown its potential in medical applications as well [25–27].

Chamomile, which has several applications in the pharmaceutical and cosmetics industries, is well known for its relaxant properties and has exhibited antioxidant activity according to a team of researchers in the Republic of Srpska, Bosnia and Herzegovina [28]. Similarly, Grapefruit, known for its use in fragrances, has also shown antioxidant activity in a study by researchers in Asia in 2010 [29].

For centuries, Ginger has been a very famous herb and is recognized by many cultures for its uses as a medicinal herb in treating digestive, respiratory, and other infections. A study in 2019 has also shown that Ginger has a significant potential use as an antifungal agent [30]. Gum Myrrh was also reported to have antifungal biological activity in a study by Perveen et al. in 2018 [31]. Ginseng has many therapeutical properties and applications as an antioxidant, antibacterial, and anticancer agent [32].

Hazelnut has medicinal traits and is recommended for mental fatigue and anemia, in addition to exhibiting antibacterial and antiparasitic activities [33,34]. Elaguel et al. showed that henna has substantial antioxidant activity and has a significant potential in combatting cancer [35]. Mint, a natural antioxidant, has also shown its effectiveness in the treatment of mental fatigue as reported in a study by a team of researchers from Korea in 2010 [29].

Mustard, a natural food preservative, has exhibited effective antimicrobial activities in a study by researchers in Iran in 2019 [36], while Egyptian researchers in 2015 have shown that onion, in addition to its many uses in the food, cosmetics, and medicinal industries, have exhibited antimicrobial and antioxidant activities [37].

Rosemary, a popular herb in the Mediterranean region especially in the food industry, has medicinal uses as well. It has been known to alleviate symptoms of respiratory and anxiety-related disorders, as well as other infectious diseases. Jiang et al. reported that Rosemary has significant antimicrobial activities [38]. In 2022, Shahrajabian examined the medicinal advantages of Sadab (Rue) and reported its many benefits as an anti-inflammatory, anti-hyperglycemic, and anti-hyperlipidemic among others [39,40]. In 2012, researchers from China and Japan examined Sandalwood for its biologically active components and reported its antioxidant, antitumor, and antitumor properties [41].

Sweet Violet is known for its therapeutic properties and for its use in the production of perfumes. It was also shown to have antioxidant and antibacterial activities as per a study conducted by Akhbari et al. in 2011 [42]. More recently, researchers in 2023 examined Turmeric given its high nutritional, industrial, and medicinal values. Specifically, due to its significant use in the food industry, it is no surprise that it has exhibited high antioxidant activities [43]. Boukhaloua et al. reported that Juniper, which is widely used in medicine, has strong antimicrobial activities [44].

The work in [45] used quantitative composition-activity relationships (QCAR) machine learning-based models to identify the chemical compounds across 61 assayed essential oils exhibiting inhibitory potency against Microsporum spp. Five different machine learning...
algorithms were used, namely, logistic regression (LR), support vector machines (SVM), gradient boosting (GB), $k$-nearest neighbor ($k$NN), and random forest (RF). Random Forest was found to be the best-performing model. The study also implements data augmentation for the biological data, which are characterized by high dimensionality and scarcity. This was conducted to dynamically alter the essential oil composition mixtures, addressing the challenge of standardizing essential oil composition due to plant and extraction method variations. Data augmentation was employed to reshape unbalanced datasets, enabling statistical analysis on larger datasets, reducing overfitting, and constructing reliable models. Our study follows a similar methodology.

3. Methodology

This section discusses the employed dataset and proposed methodology. First, a detailed analysis of the curated data is provided. After that, a discussion on the various data preparation and preprocessing steps is presented. Then, our proposed methodology for using machine learning to predict the impact of the 21 essential oils on Alzheimer’s Disease progression is detailed. The overall process entails dataset feature selection, dataset preprocessing, machine learning model evaluation, and model selection.

3.1. Dataset

The collected data (Made publicly available at https://github.com/researchrepo1/EssentialOilsDataset (accessed on 30 May 2024)) used in this research work comprise 21 essential oil samples and 428 chemical compounds, which represent the chemical composition percentages of these essential oils. These oil samples, available commercially in the United Arab Emirates market, were chosen based on the reported literature that indicate their potential activity against Alzheimer’s Disease progression [20–44,46–50]. The chemical composition data were gathered from the literature. Each essential oil sample was classified as “HIGH” or “LOW” in activity, with eight samples classified as “HIGH” and 13 as “LOW”. This classification was based directly on the literature available about each oil. Oils classified as “HIGH” reported significant activity relevant to Alzheimer’s Disease, such as neuroprotection or acetylcholinesterase inhibition. Oils classified as “LOW” either showed low activity, or there was insufficient literature to conclusively determine their efficacy.

The 21 essential oils examined in this study contain 428 chemical compounds in total, as reported in the literature. Out of which, 12 compounds were present in four or more essential oils. These compounds, along with their concentrations, are shown in Figures 1–3, which portray a visual representation of how each compound is concentrated in each oil.

As per Figure 1, the compound 1,8-Cineole is concentrated the highest in Rosemary oil at 26.54%, and is found with much lower concentrations in the remaining oils: Sweet Violet at 1.92%, Costus at 1.73%, Hazelnut at 1%, and in negligible concentrations with the remaining oils. The compound Camphene is concentrated the highest in Ginger oil at 32.79%, followed by Rosemary oil at 11.38% and Costus oil at 4.96%. It was found in less than 1% concentrations in the remaining oils. The compound Camphor is mostly concentrated in Rosemary oil at 12.88%, in Costus oil at 2.11%, and in Sweet Violet Oil at 0.92% and in Henna at 0.27%. It is noticeably clear that Limonene is highly concentrated in Grapefruit oil at 94.20%, in Juniper oil at 12.10%, and at much lower concentrations in the remaining oils. The compound Linalool is concentrated the highest in Apricot oil at 6.38%, in Sweet Violet oil at 3.06%, in Mint oil at 2.22%, in Henna oil at 1.58% and in less than 1% concentrations in the remaining oils. The compound Spathulenol is concentrated the highest in Sweet Violet oil at 2.54% and in Hazelnut oil at 1.80%. It is also found in Chamomile and Anjeer oils at 0.20% and 0.10% concentrations, respectively.

In Figure 2, it is noticed that the compound $\alpha$-Pinene is concentrated the highest in Juniper oil at 29.10%, followed closely by Rosemary oil at 20.14% and in Ginger oil at 18.05%. It is also found in less than 2% concentrations in the remaining oils. The compound $\alpha$-Terpineol is concentrated the highest in Hazelnut oil at 2.30%, in Rosemary oil at 1.95%, in
Mint oil at 0.23% and in Costus oil at 0.11%, while the compound \(\alpha\)-Thujene is concentrated the highest in Juniper oil at 2.30%, and in smaller concentrations in Rosemary, Grapefruit, Chamomile oils at 0.27%, 0.24% and 0.20%, respectively. The compound \(\beta\)-Elemene is concentrated the highest in Gum Myrrh oil at 2.20%, followed by Ginsing oil at 1.50%, at 0.40% concentration in Chamomile oil, and at a negligible 0.04% concentration in Costus oil. The compound \(\beta\)-Pinene is primarily concentrated in Juniper oil at 17.60% and at a much lower concentration of 6.59% in Rosemary oil. It is also found in several other oils but at less than 3% concentration. Finally, the compound \(\gamma\)-Cadinene is concentrated the highest in Gum Myrrh at 2.30%, in Apricot oil at 1.62%, and in both Chamomile and Juniper oils at 0.10%.

By examining the heatmap in Figure 3, it is evident that the concentration of the compound Limonene in Grapefruit oil at 94.2% is the highest among the concentrations of all the chemical compounds in the 21 oils. This is followed by the concentration of the compound Camphene in Ginger oil at 32.79%. It is worth noting that among the highest concentrations is the concentration of the compound 1,8-Cineole in Rosemary oil at 26.54%, and the concentration of the compound \(\alpha\)-Pinene in Juniper oil at 29.1%, in Rosemary oil at 20.14% and in Ginger oil at 18.05%. Furthermore, Chamomile, Costus and Rosemary oils contain nine out of the twelve compounds examined in the study. On the other hand, these 12 compounds have no concentrations in the following oils: Amla, Mustard, Onion, Sadab, Sandal, and Turmeric.

**Figure 1.** Concentrations of 1,8-Cineole, Camphene, Camphor, Limonene, Linalool, and Spathulenol compounds in essential oils.
Figure 2. Concentration s of α-Pinene, α-Terpineol, α-Thujene, β-Elemene, β-Pinene, and γ-Cadinene compounds in essential oils.

Figure 3. Heatmap representing concentrations of compounds in essential oils.

Since our dataset contains 428 chemical compounds in total, Table 1 provides a detailed overview of the concentrations of only some chemical compounds found in the 21 essential oils at hand, along with their activity label (“HIGH” or “LOW”) against the Alzheimer’s Disease progression. Each row represents an essential oil sample, and the columns list
the percentage concentrations of various compounds, including Camphene, Limonene, Linalool, α-Pinene, β-Pinene, and 1,8-Cineole. The last column labels the activity of each essential oil, as either “HIGH” or “LOW”, where “High” indicates reported significant activity of the oil and “LOW” indicates low activity or not enough literature availability about the essential oil.

**Table 1.** Data representation of 21 essential oils and only some of their chemical compositions in percentages.

<table>
<thead>
<tr>
<th>Essential Oil</th>
<th>Camphene</th>
<th>Limonene</th>
<th>Linalool</th>
<th>α-Pinene</th>
<th>β-Pinene</th>
<th>1,8-Cineole</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amla</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>LOW</td>
</tr>
<tr>
<td>Anjeer</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.6</td>
<td>0.2</td>
<td>0</td>
<td>HIGH</td>
</tr>
<tr>
<td>Apricot</td>
<td>0</td>
<td>2.54</td>
<td>6.38</td>
<td>1.37</td>
<td>0</td>
<td>0</td>
<td>HIGH</td>
</tr>
<tr>
<td>Ávocado</td>
<td>0</td>
<td>0.1</td>
<td>0.01</td>
<td>0.67</td>
<td>0</td>
<td>0.3</td>
<td>LOW</td>
</tr>
<tr>
<td>Chamomile</td>
<td>0</td>
<td>0.3</td>
<td>0.3</td>
<td>1.9</td>
<td>0.1</td>
<td>0.1</td>
<td>HIGH</td>
</tr>
<tr>
<td>Costus</td>
<td>4.96</td>
<td>0.51</td>
<td>0.4</td>
<td>1.02</td>
<td>0.04</td>
<td>1.73</td>
<td>HIGH</td>
</tr>
<tr>
<td>Ginger</td>
<td>32.79</td>
<td>0</td>
<td>0.59</td>
<td>18.05</td>
<td>2.95</td>
<td>0</td>
<td>HIGH</td>
</tr>
<tr>
<td>Ginseng</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>LOW</td>
</tr>
<tr>
<td>Grapefruit</td>
<td>0</td>
<td>94.2</td>
<td>0</td>
<td>1</td>
<td>2.17</td>
<td>0</td>
<td>LOW</td>
</tr>
<tr>
<td>Gum Myrrh</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>LOW</td>
</tr>
<tr>
<td>Hazelnut</td>
<td>0</td>
<td>3.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>LOW</td>
</tr>
<tr>
<td>Henna</td>
<td>0.13</td>
<td>0</td>
<td>1.58</td>
<td>0.16</td>
<td>0.54</td>
<td>0</td>
<td>LOW</td>
</tr>
<tr>
<td>Juniper</td>
<td>0.9</td>
<td>12.1</td>
<td>0</td>
<td>29.1</td>
<td>17.6</td>
<td>0</td>
<td>LOW</td>
</tr>
<tr>
<td>Mint</td>
<td>0</td>
<td>0</td>
<td>2.22</td>
<td>0.69</td>
<td>1.14</td>
<td>0</td>
<td>HIGH</td>
</tr>
<tr>
<td>Mustard</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>LOW</td>
</tr>
<tr>
<td>Onion</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>LOW</td>
</tr>
<tr>
<td>Rosemary</td>
<td>11.38</td>
<td>1.32</td>
<td>0.25</td>
<td>20.14</td>
<td>6.95</td>
<td>26.54</td>
<td>HIGH</td>
</tr>
<tr>
<td>Sadab</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>LOW</td>
</tr>
<tr>
<td>Sandal</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>HIGH</td>
</tr>
<tr>
<td>Sweet violet</td>
<td>0</td>
<td>0</td>
<td>3.06</td>
<td>1.31</td>
<td>0.62</td>
<td>1.92</td>
<td>LOW</td>
</tr>
<tr>
<td>Turmeric</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>LOW</td>
</tr>
</tbody>
</table>

3.2. Feature Selection

As reported in Table 1, our dataset has more than 400 attributes. This data phenomenon is known as the curse of high-dimensionality, whereby data points are sparse in the high-dimensional space, resulting in a negative impact on the predictive accuracy of machine learning models [51].

Due to the extremely high dimensionality of our dataset, a rigorous feature selection process is implemented to enhance the predictive accuracy of the machine learning models. This process entails creating various versions of the dataset by systematically excluding chemical compound features that exhibit relatively low cumulative percentages across all essential oil samples.

Each chemical compound could theoretically achieve a maximum percentage of 100% within a single essential oil, yielding a total possible sum of 2100% across all 21 essential oils. In practice, the highest observed cumulative percentage was for Limonene, with a value of 114.97%. To determine which features to retain, percentage thresholds were established based on this maximum value. 11 distinct threshold values, particularly 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 30, 40, and 50, were enforced, resulting in 11 dataset variations (subsets). Additionally, the original dataset, incorporating all chemical compound features (where the sum is greater than 0), is also included for comparison. This approach results in twelve unique datasets, each progressively refining the feature set to enhance the model’s efficacy.
in predicting the activity of essential oils against Alzheimer’s Disease progression, from a computational perspective.

Table 2 specifies the 11 threshold values enforced in the feature selection process, the corresponding number of resulting chemical compound features, and the designated names for each dataset. The experiments in Section 4 are carried out on all the 11 + 1 (original) datasets.

Table 2. Composition distribution of the twelve datasets.

<table>
<thead>
<tr>
<th>Dataset Number</th>
<th>Composition Sum Threshold Criteria</th>
<th>Number of Chemical Compounds</th>
<th>Dataset Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sum greater than 0</td>
<td>428</td>
<td>all_comounds.csv</td>
</tr>
<tr>
<td>2</td>
<td>Sum greater than 0.1</td>
<td>416</td>
<td>more_than_0.1_comounds.csv</td>
</tr>
<tr>
<td>3</td>
<td>Sum greater than 0.2</td>
<td>389</td>
<td>more_than_0.2_comounds.csv</td>
</tr>
<tr>
<td>4</td>
<td>Sum greater than 0.5</td>
<td>301</td>
<td>more_than_0.5_comounds.csv</td>
</tr>
<tr>
<td>5</td>
<td>Sum greater than 1</td>
<td>226</td>
<td>more_than_1_comounds.csv</td>
</tr>
<tr>
<td>6</td>
<td>Sum greater than 2</td>
<td>139</td>
<td>more_than_2_comounds.csv</td>
</tr>
<tr>
<td>7</td>
<td>Sum greater than 5</td>
<td>77</td>
<td>more_than_5_comounds.csv</td>
</tr>
<tr>
<td>8</td>
<td>Sum greater than 10</td>
<td>46</td>
<td>more_than_10_comounds.csv</td>
</tr>
<tr>
<td>9</td>
<td>Sum greater than 20</td>
<td>25</td>
<td>more_than_20_comounds.csv</td>
</tr>
<tr>
<td>10</td>
<td>Sum greater than 30</td>
<td>16</td>
<td>more_than_30_comounds.csv</td>
</tr>
<tr>
<td>11</td>
<td>Sum greater than 40</td>
<td>9</td>
<td>more_than_40_comounds.csv</td>
</tr>
<tr>
<td>12</td>
<td>Sum greater than 50</td>
<td>8</td>
<td>more_than_50_comounds.csv</td>
</tr>
</tbody>
</table>

3.3. Preprocessing

Machine learning algorithms may require numerical data in specific formats for optimal performance. Initially, our dataset contains textual labels for activity and unformatted numerical data for features. To facilitate data processing, the binary classes “HIGH” and “LOW” are converted to numerical values using Label Encoding. Label Encoding is a technique that transforms categorical text data into numerical values. In this case, the labels “HIGH” and “LOW” are encoded as 1 and 0, respectively.

Furthermore, machine learning algorithms generally perform better when the numerical values are normalized within a specific range, typically between 0 and 1. However, the values in our datasets ranged from 0 to 100, representing percentage compositions. To address this issue, the Min-Max Scaling technique is employed. Min-Max Scaling transforms the data by rescaling each feature to a specified range, usually 0 to 1. This is achieved by subtracting the minimum value of the feature from each data point and then dividing by the range of the feature (the difference between the maximum and minimum values), as per Equation (1).

$$X' = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

In Equation (1), $X$ is the original value, $X_{\text{min}}$ is the minimum value of the feature, $X_{\text{max}}$ is the maximum value of the feature, and $X'$ is the scaled value. This normalization process ensures that the selected machine learning models are effectively trained, as the dataset now contains standardized feature values to a common scale.

Further preprocessing beyond these steps was not necessary, as the dataset does not contain any outliers. All numerical values are between 0 and 100. Moreover, the dataset does not contain any missing values as all chemical compounds were accounted for during the process of collecting their data from the literature.
3.4. Model Training

The subsequent step in developing a machine learning model involves splitting the dataset into training and testing sets. Traditionally, the training set comprises 70% of the samples, with the remaining 30% allocated to the testing set [52]. This approach is generally more effective for larger datasets containing at least 100 samples. Given that the current dataset consists of only 21 samples, a 50%–50% split with stratification is employed.

Stratification is particularly important due to the imbalanced nature of the original dataset, which has a ratio of 8:13 for “HIGH” to “LOW” activity labels. Stratification ensures that each subset (training and testing) maintains a distribution of class labels close to the original dataset. By employing stratified sampling, the ratio of “HIGH” to “LOW” labels is preserved within both the training and testing sets. This ensures that the model is exposed to a representative distribution of the classes, enhancing prediction accuracy in classification tasks.

3.5. Model Evaluation

The effectiveness of a machine learning model is evaluated based on its accuracy in predicting the correct label of an essential oil sample in the underlying dataset (see Tables 1 and 2). Particularly, two accuracy metrics are employed: testing accuracy and Leave-one-out cross-validation (LOOCV) accuracy.

Cross-validation (CV) is a model evaluation technique where the sample set is split into \( k \) subsets, and model fitting and prediction are performed \( k \) times. The testing set will be formed of one particular subset, out of the \( k \) subsets, and the remaining subsets will be the training set. This method is useful for avoiding model overfitting and also for effectively evaluating the performance of a model on a smaller dataset.

Leave-one-out cross-validation (LOOCV) is a form of \( k \)-fold cross-validation, where \( k = n \) and \( n \) is the number of samples in the dataset. In the \( n \) iterations, the testing set will be the \( k \)th sample and the remaining samples will be the training set. To evaluate model performance using LOOCV, the accuracies of the model in predicting the activity of the testing sample in the \( n \) training iterations are averaged to give the overall LOOCV accuracy of the model.

Due to the extremely small size of our dataset in terms of the number of samples, several other evaluation methods were experimented with to identify the best ones. These evaluation methods include 50:25:25 and 70:15:15 ratios of training-validation-testing sets, and mean cross-validation using \( k = 3, 5, 7 \), and LOOCV. Out of these methods, evaluating a model using 50% training and 50% testing sets and using LOOCV reported the highest accuracies. Table 3 summarizes these experiments. Henceforth, all our experiments will be evaluated using the 50% testing set and LOOCV. The accuracies of a machine learning model in predicting the activity of essential oils are hereby referred to as testing accuracy and LOOCV accuracy, respectively.

Table 3. Average accuracy for each algorithm using different model evaluation techniques.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN</td>
<td>36.88</td>
<td>35.72</td>
<td>78.03</td>
<td>32.42</td>
<td>30.5</td>
<td>59.92</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>60.42</td>
<td>40.27</td>
<td>86.11</td>
<td>33.47</td>
<td>33.5</td>
<td>64.68</td>
</tr>
<tr>
<td>Random Forest</td>
<td>38.03</td>
<td>39.29</td>
<td>70.45</td>
<td>39.26</td>
<td>39.29</td>
<td>67.06</td>
</tr>
</tbody>
</table>

3.6. Model Selection

The problem of predicting the class label of an essential oil sample is a classification problem. This section discusses building three widely-used classification models: \( k \)-Nearest Neighbours, Logistic Regression, and Random Forest. For each one of these models, this section discusses the selection of the various model parameters. All these algorithms were implemented using Python’s scikit-learn library [45,53,54].
3.6.1. K-Nearest Neighbours

K-Nearest Neighbors (kNN) is a supervised learning classification algorithm that operates by plotting samples in a multidimensional space, where the number of features corresponds to the number of dimensions. The class label of a sample is predicted by identifying the k nearest neighbors in this multidimensional space and taking a majority vote of these neighbors’ labels [55].

It is crucial to select an appropriate value for k to avoid ties in the voting process. For binary classification tasks, k should be an odd number to prevent ties. In cases where the classification involves three classes, k should be chosen such that it is not a multiple of three, minimizing the likelihood of tie votes. This careful selection of k ensures accurate classification outcomes.

To explore the impact of different parameters on the performance of the kNN algorithm, the scikit-learn implementation was utilized [53]. Specifically, an odd number of neighbors (the n_neighbors parameter) is chosen between 1 and 9, and two distinct settings for the weights parameter, uniform and distance, are used. The uniform setting assigns equal weights to each neighbor, while the distance setting assigns larger weights to closer samples [53]. This configuration yields 10 distinct kNN models, each characterized by a combination of parameters. These models are listed and named in Table 4.

Table 4. kNN models and their parameters.

<table>
<thead>
<tr>
<th>Weights</th>
<th>k = 1</th>
<th>k = 3</th>
<th>k = 5</th>
<th>k = 7</th>
<th>k = 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>kNN_M1</td>
<td>kNN_M3</td>
<td>kNN_M5</td>
<td>kNN_M7</td>
<td>kNN_M9</td>
</tr>
<tr>
<td>Distance</td>
<td>kNN_M2</td>
<td>kNN_M4</td>
<td>kNN_M6</td>
<td>kNN_M8</td>
<td>kNN_M10</td>
</tr>
</tbody>
</table>

3.6.2. Logistic Regression

Logistic Regression (LR) is a supervised learning binary classification algorithm that predicts the probability of a sample belonging to a certain class. Equation (2) formulates this process, where the equation takes all input features as weighted variables, and the output would be either 0 or 1 with each extremity representing the two classes [56].

\[ y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_n x_n \]  

Equation (2)

In Equation (2), y is either 0 or 1, the input features are \( x_1, x_2, x_3, \ldots, x_n \), and the weights of the features are given by \( w_1, w_2, w_3, \ldots, w_n \), where n is the number of features.

The output can be plotted as an S-shaped curve, which is also called a Sigmoid function. The curve helps to visualize the predicted probabilities and how outputs are assigned. This model is called regression since the prediction is a range of values between 0 and 1, but it performs binary classification since the output can only have two possible values, 0 or 1.

During the model training process, the weights are calculated based on the output values. Then, during model evaluation, the testing samples are assigned output values by substituting the values of their input features into the equation, multiplying each of them with their respective calculated weights, adding them up, and the resulting output would be assigned 0 or 1 based on the value to which it is closest.

An optimization algorithm is used to find the optimal weights and bias during training. Conventionally, gradient descent is used for this purpose. An LR Model trained on the same dataset with different optimization algorithms may result in different outcomes.

To prevent overfitting the S-shaped curve to the training data, there are regularization methods called L1 Loss and L2 Loss functions. An LR model can be trained using either one of these loss functions or neither of them. Implementing a loss function improves model performance by penalizing incorrect predictions.

Using Python’s scikit-learn library, we experimented with several parameters for Logistic Regression. The penalty parameter trains the LR model with either L1 or L2 loss functions (or none) to regularize the data and avoid overfitting to the training set. The
The `class_weight` parameter is `None` by default but can be set to `balanced` for an imbalanced training set. This assigns samples of the minority class with a higher weight and majority class samples with a lower weight. This may improve model performance. The `solver` parameter implements different optimization algorithms. The `C` parameter specifies the strength of regularization, whereas a smaller `C` would specify stronger regularization [54].

We experimented with the `penalty`, `solver`, and `C` parameters and found that `class_weight = 'balanced'` and `penalty = 'L2'` worked well for all datasets. We then varied `C` and `solver`, and evaluated the performance of the logistic regression models. Table 5 summarizes the different LR models resulting from parameter variations.

**Table 5. Logistic Regression models and their parameters.**

<table>
<thead>
<tr>
<th>Solver/C</th>
<th>C = 0.001</th>
<th>C = 0.01</th>
<th>C = 0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>lbfgs</td>
<td>LR_M1</td>
<td>LR_M5</td>
<td>LR_M9</td>
</tr>
<tr>
<td>liblinear</td>
<td>LR_M2</td>
<td>LR_M6</td>
<td>LR_M10</td>
</tr>
<tr>
<td>newton-cg</td>
<td>LR_M3</td>
<td>LR_M7</td>
<td>LR_M11</td>
</tr>
<tr>
<td>newton-cholesky</td>
<td>LR_M4</td>
<td>LR_M8</td>
<td>LR_M12</td>
</tr>
</tbody>
</table>

### 3.6.3. Random Forest

Random Forest (RF) is an ensemble supervised learning technique that can be used for classification and regression applications [57]. To understand how RF works, one must first understand how Decision Trees (DT) works. Decision Trees is another machine learning technique that is used for classification and regression. It works as a flowchart, splitting the dataset based on different features at different levels, to ultimately predict the label of a target sample [58].

Random Forest implements multiple decision trees, where a random subset of features is dropped in each decision tree, thus minimizing the chance of overfitting to the training data. The class label of a sample is predicted by feeding that sample’s feature values into each decision tree, obtaining each decision tree’s predicted class label, and then taking the majority of votes to assign the output class label [59].

An RF model’s accuracy is based on multiple parameters, namely the number of decision trees, the best-split algorithm, and whether the tree has been pre-pruned or not. Pre-pruning refers to limiting the growth of the trees. This is conducted to further ensure that the model does not overfit. One way of conducting this is by limiting the maximum depth of the trees or by varying certain parameters during training.

The `n_estimators` parameter refers to the number of decision trees used in the random forest. The `criterion` parameter specifies the method of choosing the best feature for splitting the data at each node. We experimented with the `gini`, `log_loss`, and `cross-entropy` methods for `criterion` and discovered that the `log_loss` method results in a better performance. The `max_depth` and `ccp_alphas` parameters are used to pre-prune the trees. `max_depth` refers to the maximum depth of each decision tree in the RF model, whereas `ccp_alphas` is a constant that specifies which sub-trees are allowed in a singular tree based on their cost complexity. By limiting the depth and complexity of the trees, we can ensure better performance. Table 6 summarizes the different RF models resulting from parameter variations.

**Table 6. Random Forest models and their respective parameters.**

<table>
<thead>
<tr>
<th>Varying Parameter</th>
<th>n_estimators</th>
<th>criterion</th>
<th>max_depth</th>
<th>ccp_alphas</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF_1</td>
<td>n_estimators</td>
<td>1–20</td>
<td>log_loss</td>
<td>6</td>
</tr>
<tr>
<td>RF_2</td>
<td>max_depth</td>
<td>4</td>
<td>log_loss</td>
<td>2–14</td>
</tr>
<tr>
<td>RF_3</td>
<td>ccp_alphas</td>
<td>4</td>
<td>log_loss</td>
<td>N/A</td>
</tr>
</tbody>
</table>
4. Results

In this section, we carry out extensive experiments using the three classification methods detailed in Section 3.6 and datasets in Table 2. All the classification models are implemented using Python’s scikit-learn library [45,53,54]. The performance of a model is evaluated based on its prediction accuracy using two metrics: testing accuracy and LOOCV accuracy.

Certain models exhibited a notable trend wherein performance seemed to improve across the last five datasets (Table 2). These datasets, characterized by cumulative percentage compositions of each chemical compound exceeding 10%, were subject to focused examination. In visualizing the outcomes of these experiments, the average accuracy across all twelve datasets is computed. Additionally, to provide nuanced insights, the results from the last five datasets are visualized separately. This approach helps in identifying datasets that demonstrate better performance across all models.

4.1. k-Nearest Neighbors Results

The ten $k$NN models specified in Table 4 were evaluated by plotting the varying parameters with testing and LOOCV accuracies. Figure 4 describes the performance of these $k$NN models.

![Graph showing performance of kNN models](image)

From Figure 4a,b, it can be seen that the best-performing model is $kNN_M4$. With reference to Table 4, this result suggests that $k = 3$ and `weights = 'distance'` are the best
model parameters. It is also worth noting that the testing accuracies are higher than the LOOCV accuracies. This may indicate a bias in our training set, despite stratification.

4.2. Logistic Regression Results

The twelve LR models specified in Table 5 are evaluated by plotting their testing and LOOCV accuracies. Figure 5 shows the accuracies of all the models averaged across all the twelve datasets, and the accuracies across the last five datasets only.

![Logistic Regression Model](image)

**Figure 5.** Average testing and LOOCV accuracies for the Logistic Regression models. (a) Across all the twelve datasets; (b) Across the last five datasets.

Figure 5a,b report that model LR_M10 seems to perform the best. With reference to Table 5, \( C = 0.1 \), \( \text{solver} = \text{‘liblinear’} \), \( \text{penalty} = \text{L2} \), and \( \text{class_weight} = \text{‘balanced’} \) are the best model parameters. It is also worth noting that there is a difference between the testing and LOOCV accuracies, and this may indicate a bias in our training data. It is also likely that the LR models are not performing well as the chemical composition data may not be linearly related.

4.3. Random Forest Results

Table 6 summarizes three RF models, RF_1, RF_2, and RF_3, by varying the \( n_{\text{estimators}} \), \( \text{max\_depth} \), and \( \text{ccp\_alphas} \) parameters. That is, for each model, one parameter is varied while the other two are fixed. The accuracy results against testing and LOOCV metrics of each RF model are plotted twice, once using all the twelve datasets and another using only the last five datasets. These results are plotted in Figure 6, Figure 7 and Figure 8, for each model, respectively.
Figure 6. Average testing and LOOCV accuracies for the n_estimators parameter. (a) Across all the twelve datasets; (b) Across the last five datasets.

Figure 6a,b report the performance of model RF_1. Both figures show that the graphs peak at n_estimators = 6 and 12, suggesting that these two parameter values result in the model having the highest accuracy, while the other two parameters are fixed.

Figure 7. Average testing and LOOCV accuracies for the max_depth parameter. (a) Across all the twelve datasets; (b) Across the last five datasets.

Looking at Figure 7a,b, which plot the performance of model RF_2, we can see the disparate performance of the same model against the two evaluation metrics, testing and LOOCV. LOOCV results in much better prediction accuracy than testing, where the model performance peaks around max_depth = 4 for LOOCV.

Evaluating the performance of model RF_3 in Figure 8a,b, the accuracy is highest at ccp_alphas = 0.25 for the LOOCV metric.

Based on Figures 6–8, we can conclude that the best model parameters for the RF model would be n_estimators = 12, max_depth = 4, and ccp_alphas = 0.25.

All the experiments conducted thus far used the datasets whose dimensionality was reduced (from 428 dimensions) as per the feature selection process presented in Section 3.2. To maintain a more objective evaluation approach, we employ Principal Component Analysis (PCA) for dimensionality reduction [60,61]. PCA transforms a high-dimensional dataset by projecting data points onto a new subspace defined by newly created dimensions, i.e., principal components. The idea is to utilize the original dimensions that contribute the most to explaining the data. The latter is conducted through the variance. That is, we perform PCA on our curated dataset (see Section 3.1) where the principal components explain 90% and 95% of the variance, respectively. This leads to the creation of two datasets; the first dataset contains 16 principal components, and the second dataset contains 18 principal components, respectively.
In Figure 9, PCA was conducted such that 90% of the data variance is maintained. This resulted in a dataset that contains 16 dimensions. We observe from both Figure 9a,b that the prediction model exhibits overfitting when trained and tested using a 50%–50% split. As such, it would not be fair to accept or conclude the high accuracy of 90.91% reported in Figure 9b.

Figure 8. Average testing and LOOCV accuracies for the ccp_alphas parameter. (a) Across all the twelve datasets; (b) Across the last five datasets.

Figure 9. Model accuracies after PCA with 16 principal components explaining 90% of the variance. (a) kNN; (b) Logistic Regression.

Similarly, Figure 10 reports the prediction results from using PCA where 95% of the data variance is explained by 18 principal components (dimensions). Figure 10a depicts
fluctuation of accuracy across the different kNN models (see Table 4), but the model stabilizes under the configuration of “kNN_M9” and “kNN_M10”. On the other hand, Figure 10b shows a robust and stable performance across most of the Logistic Regression models (see Table 5) and far fewer overfitting issues. Consequently, under “LR_M11”, the highest reported accuracy can be considered.

Figure 10. Model accuracies after PCA with 18 principal components explaining 95% of the variance. (a) kNN; (b) Logistic Regression.

5. Discussion

This section discusses the results obtained from our experiments and adds context to their applicability in clinical studies.

5.1. Experimental Results

In Section 4, we extensively experimented with three prominent classification models, kNN, LR, and RF, by varying their respective parameters and underlying datasets. Table 7 summarizes the parameters under which each model performs the best across all twelve datasets (Table 2).

Next, we compare the accuracy results from using each of the twelve datasets from Table 2 in order to determine which dataset performs the best overall. The winner dataset will be used in our last experiment later in this section. We apply the best-performing kNN, LR, and RF models to each of the twelve datasets, then we take the average accuracy across the three models.
Table 7. The specifications of the best kNN, LR, and RF models.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Model Parameters</th>
<th>Best Performing Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN</td>
<td>$k$</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>weights distance</td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>penalty</td>
<td>L2</td>
</tr>
<tr>
<td></td>
<td>class_weight</td>
<td>balanced</td>
</tr>
<tr>
<td></td>
<td>solver</td>
<td>liblinear</td>
</tr>
<tr>
<td></td>
<td>$C$</td>
<td>0.1</td>
</tr>
<tr>
<td>Random Forest</td>
<td>criterion</td>
<td>log_loss</td>
</tr>
<tr>
<td></td>
<td>n_estimators</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>max_depth</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>ccp_alphas</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Figure 11 plots the average performance of the three best models against each of the twelve datasets. The average testing and LOOCV accuracies resulting from each dataset are reported. Comparing all twelve datasets, more_than_50_compounds results in the highest accuracy. It is also worth noting that, in general, the average LOOCV accuracy across the different datasets is higher than the average testing accuracy. This observation has been consistent in all the Logistic Regression and Random Forest experiments depicted in Figures 5–8.

Figure 11. Average performance of the three best algorithms across the different datasets.

To determine which of the three models performs best, we apply the best kNN, LR, and RF models from Table 7 on the best dataset determined in Figure 11, i.e., more_than_50_compounds. Figure 12 depicts this comparison. The Logistic Regression and Random Forest models having the parameters listed in Table 7 result in the highest accuracy of nearly 81% on the LOOCV metric, while kNN leads on the testing metric with an accuracy of nearly 82%. This conclusion corroborates our observations from the literature [62].

Due to the large number of features in our original dataset (see Section 3.1), one key aspect of this study is dimensionality reduction. Feature selection was conducted following our proposed method described in Section 3.2, and using Principal Component Analysis (PCA) [60,61]. Both dimensionality reduction methods resulted in similar model performances, achieving around 81% accuracy. Figure 12 also depicts the best result achieved by preprocessing the dataset using PCA.
5.2. Application

This study harnesses machine learning to explore the potential therapeutic properties of essential oils in treating Alzheimer’s Disease. Our study underscores the potential of integrating computational models with traditional pharmacological approaches to accelerate the discovery of novel therapeutic agents.

The biological relevance of selected compounds provides insights into possible mechanisms of action against Alzheimer’s Disease progression. 1,8-Cineole, commonly found in eucalyptus oil, has been documented for its anti-inflammatory and antioxidant properties [63]. Research indicates that 1,8-Cineole can modulate the activity of key enzymes and inflammatory mediators in the brain, which are crucial in the pathology of Alzheimer’s Disease. By potentially reducing oxidative stress, 1,8-Cineole is reported for its significant antioxidant and anti-Alzheimer’s Disease activities, which provide a potential medicinal approach for Alzheimer’s Disease [64].

Camphene, another significant compound identified, exhibits strong antioxidant properties that may protect neuronal cells from oxidative stress which is a known contributor to Alzheimer’s Disease. Found in high concentrations in ginger and rosemary oils, Camphene’s role in reducing oxidative damage underscores its relevance in slowing disease progression [65].

Additionally, α-Pinene has demonstrated potential in inhibiting acetylcholinesterase, an enzyme associated with the degradation of the neurotransmitter acetylcholine, which is notably diminished in Alzheimer’s Disease patients. By modulating acetylcholine levels, α-Pinene may improve cognitive function and communication between neurons, providing a plausible mechanism through which essential oils could benefit Alzheimer’s Disease patients [66].

Understanding the interactions of these bioactive compounds with the biological pathways affected by Alzheimer’s Disease allows us to hypothesize their mechanisms of action. This not only enhances our understanding of disease pathology but also directs future research toward interventions that could modulate these pathways more effectively. For instance, the anti-inflammatory properties of 1,8-Cineole could be leveraged to develop treatments that target inflammation-related pathways in Alzheimer’s Disease.

The insights gained from our experimental analysis suggest new directions for research into the therapeutic potential of essential oils and their constituents. Further in vivo and clinical studies are necessary to validate these findings and to assess the efficacy and safety of these compounds in human populations.

Figure 12. Visualizing the performance of the three classification models on the more_than_50_compounds and PCA datasets.
This study proposes a framework for integrating essential oils into existing treatment strategies. Compounds such as 1,8-Cineole, Camphene, and α-Pinene, highlighted for their neuroprotective and anti-inflammatory properties, could synergistically enhance the effectiveness of current pharmacological treatments like cholinesterase inhibitors and NMDA receptor antagonists. Moreover, the antioxidative properties of these compounds suggest their use in preventative strategies aimed at high-risk populations, potentially delaying or even preventing the onset of Alzheimer’s symptoms. This study invites a re-evaluation of current treatment protocols and encourages the development of holistic, multi-targeted treatment approaches. These insights could foster innovative clinical trials and might lead to the development of a new class of neuroprotective medications that address the complex pathophysiology of Alzheimer’s Disease more comprehensively.

6. Conclusions and Future Work

This study was conducted in the hope of bridging the gap between traditional knowledge of essential oils and modern machine learning techniques and offering a unique perspective on Alzheimer’s Disease therapeutics. Particularly, this study investigates the predictive capability of machine learning techniques on the impact of plants’ essential oils on Alzheimer’s Disease progression.

We curated data from 21 essential oils commercially available in the United Arab Emirates market. The 21 essential oil samples have a total of 428 chemical compounds, which, from a data analytics perspective, result in an extremely high-dimensional dataset that renders predictive models useless due to data sparsity. Extensive data preprocessing and feature selection steps were carried out in order to increase the efficacy of the applied machine learning models.

Several classification models were configured and evaluated against two accuracy metrics. Experimental results showcased promising accuracy in predicting essential oils activity against Alzheimer’s Disease progression, and suggest that Random Forest and Logistic Regression have the potential to be nearly 82% accurate.

We emphasize that our findings require further in vivo and clinical studies are necessary to validate these findings and to assess the efficacy and safety of these compounds in human populations. This is an important and natural progression to this work.

While our models exhibited robust performance, the exploration must continue. One potential future work is to validate the findings of this study clinically to measure the effectiveness and safety of essential oil compounds in human populations. Another future work is to employ and discuss more feature selection and model validation techniques and compare their impact on prediction capability. A third future direction would be employing data augmentation techniques and integrating neural networks to unravel complex relationships within chemical features. This holistic approach aims to enhance model generalization and uncover nuanced patterns, contributing to a more comprehensive understanding of essential oils’ efficacy in preventing Alzheimer’s Disease progression and exploring potential treatments.


Funding: This research is supported in part by Research Fund 2023-24-1004 from Rochester Institute of Technology, Dubai.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.
Data Availability Statement: The raw data that support the findings of this article are publicly available on GitHub at https://github.com/researchrepo1/EssentialOilsDataset (accessed on 30 May 2024).

Acknowledgments: The authors would like to thank the reviewers for their constructive feedback that contributed to the enhancement of this study.

Conflicts of Interest: The authors have no competing interest to declare that are relevant to the content of this article.

Abbreviations
The following abbreviations are used in this manuscript:

- FDA Food and Drug Administration
- CNS Central Nervous System
- SVM Support Vector Machine
- kNN k-Nearest Neighbor
- LR Logistic Regression
- RF Random Forest

References
10. Rahmi, D.; Yunilawati, R.; Jati, B.; Setiawati, I.; Ristyanto, A.; Batubara, I.; Astuti, R. Antiaging and Skin Irritation Potential of Four Main Indonesian Essential Oils. Cosmetics 2021, 8, 94. [CrossRef]