Advances in Technology Innovation, vol. 9, no. 2, 2024, pp. 85-98

Precision Geolocation of Medicinal Plants: Assessing Machine Learning Algorithms for Accuracy and Efficiency

Maria Concepcion Suarez Vera*

College of Information and Communications Technology, Catanduanes State University, Catanduanes, Philippines

Received 06 February 2024; received in revised form 08 March 2024; accepted 09 March 2024

DOI: https://doi.org/10.46604/aiti.2024.13355

Abstract

This study investigates the precision geolocation of medicinal plants, a critical endeavor bridging ecology, conservation, and pharmaceutical research. By employing machine learning algorithms—gradient boosting machine (GBM), random forest (RF), and support vector machine (SVM)—within the cross-industry standard process for data mining (CRISP-DM) framework, both the accuracy and efficiency of medicinal plant geolocation are enhanced. The assessment employs precision, recall, accuracy, and F1 score performance metrics. Results reveal that SVM and GBM algorithms exhibit superior performance, achieving an accuracy of 97.29%, with SVM showing remarkable computational efficiency. Meanwhile, despite inferior performance, RF remains competitive especially when model interpretability is required. These outcomes highlight the efficacy of SVM and GBM in medicinal plant geolocation and accentuate their potential to advance environmental research, conservation strategies, and pharmaceutical explorations. The study underscores the interdisciplinary significance of accurately geolocating medicinal plants, supporting their conservation for future pharmaceutical innovation and ecological sustainability.

Keywords: geolocation, machine learning, medicinal plants, support vector machine, gradient boosting machine

1. Introduction

Tracing back to ancient civilizations and extending into modern ecological conservation and pharmaceutical domains, the precise geolocation of medicinal plants is substantiative regarding the enhancement of healthcare outcomes, preserving biodiversity, and promoting sustainable development. Medicinal plants, integral to the healing traditions of Egyptians, Chinese, Indians, and other cultures, have been perpetually used to prevent, relieve, or treat illnesses. This practice, profoundly embedded in the cultural heritage of numerous communities, has been meticulously documented and passed down through generations. In the Philippines, the melding of Malay, Spanish, and American influences enrich its traditional understanding of medicinal plants, with conventional healers such as “albularyos” or “pilot” using these plants to treat various ailments.

Ethnobotanical research in the Philippines highlights the deep traditional knowledge of indigenous tribes, identifying the country as a critical biodiversity hotspot with around 13,000 plant species, 39% of which are endemic [1-2]. This biodiversity underpins the extensive use of 1,500 medicinal plants in traditional medicine, with significant potential recognized for contemporary pharmaceuticals. Among this vast cluster of medicinal plants, 10 plants are widely recognized, 177 are earmarked for further research, and the confirmation of safety and efficacy pertains to 120 plants [3]. These insights emphasize the significance of medicinal plants in both traditional and modern healthcare contexts, showcasing their potential in pharmaceutical development.

* Corresponding author. E-mail address: maconsuarez@gmail.com

English language proofreader: Chih-Wei Chang

The traditional geolocation methods for medicinal plants, including field surveys and basic GPS mapping, have been instrumental yet exhibit a palpable defect in accuracy, efficiency, and data integration, as substantiated in Faizy et al. [4]. Furthermore, the paucity of integration of machine learning (ML) techniques further compounds this limitation, significantly enhancing the precision and efficiency of geolocation practices. The critical role of geolocation accuracy in scientific endeavors is underscored by Halpin et al. [5], who demonstrate its influence on the retrieval of wind field data, thus affirming the necessity for refined geolocation methods across various scientific applications. Additionally, the exploration of agroecological zoning models highlights the integration of climatic and edaphic parameters to optimize the growth of medicinal plants, presenting a methodological advancement in identifying potential growth areas [6].

Despite the advent of applications employing advanced technologies such as crowdsourcing, image recognition, and convolutional neural networks for the identification and recognition of medicinal plants, as seen in Isa et al. [7] and Sugiarto et al. [8], the field remains challenged by the need for more precise and efficient geolocation methods. The utilization of geospatial database management systems and the development of augmented reality portals, as presented in Puttinaovarat and Horkaew [9] and Permana et al. [10] respectively, indicate a technological evolution to enhance interaction with medicinal plant information. Nonetheless, the enduring value of traditional knowledge, as documented in Faruque et al. [11] and the ethnobotanical analysis in Boycheva and Ivanov [12], emphasizes the integration of such a posteriori knowledge into contemporary technological advancements.

Given these aforementioned considerations, this research posits a compelling argument for adopting innovative approaches that leverage the latest technological advancements to conquer the current limitations in medicinal plant geolocation. By accentuating the enhancement of precision and efficiency of geolocation techniques, the study endeavors to make significant contributions to the fields of conservation, sustainable harvesting, and pharmaceutical development, underscoring the importance of accurately mapping plant species for the protection of biodiversity, the maintenance of ecosystem balance, and the facilitation of drug discovery processes.

ML algorithms have emerged as a scientifically pivotal innovation, providing the tools for in-depth analysis of intricate environmental and biological datasets. This advancement surpasses conventional methodologies by facilitating precise forecasts of the locations of medicinal plants, thereby improving geolocation accuracy and fostering new research opportunities.

This research aims to enhance the accuracy of geolocating medicinal plants through a thorough analysis, evaluating the effectiveness of gradient boosting machine (GBM), random forest (RF), and support vector machine (SVM) comparatively. These algorithms, each celebrated for their distinctive benefits and empirical effectiveness in various sectors, are systematically employed to address the unique challenges presented by the geolocation of medicinal plants.

GBM learning techniques have demonstrated considerable success across various domains. Researchers have applied gradient boosting in agriculture to predict crop yield, as mentioned in Anbananthen et al. [13]. Besides, researchers have employed gradient boosting in healthcare to predict adverse outcomes in pneumonia patients [14] and to forecast cardiovascular diseases [15]. These applications underscore gradient boosting's effectiveness in achieving high prediction accuracy. Moreover, gradient boosting has shown superior accuracy and prediction performance compared with other ML algorithms, like deep learning and RF [16].

RF has undergone thorough investigation and found wide application across multiple fields, such as agriculture and healthcare, attributed to its reliability and efficiency in predictive modeling. Researchers have deployed RF to predict crop yields [17] and classify agriculture farm machinery [18]. These studies have demonstrated the high accuracy and precision of RF in agricultural applications, being acknowledged as valuable tools for decision support in farming and crop management. In healthcare, the RF has shown promising results in various applications, such as predicting the severity of patient falls [19] and forecasting hospital readmissions [20].

These research findings highlight the efficacy of RF by showcasing metrics such as accuracy, precision, recall, and F1 score in a high-performing context, accentuating its value in predictive modeling and decision-making within the healthcare sector.

SVM, rooted in statistical learning theory, is renowned for its strong performance in diverse fields. Its effectiveness spans a range of applications, from stench detection to yield prediction and complex tasks in computer science, especially spam comment screening on YouTube and real-time emotion detection [21-23]. Similarly, Shi et al. [24] utilized SVM for crop yield prediction in agriculture, achieving high accuracy. Furthermore, as substantiated by Suresh et al. [25], it is found that SVM outperformed other ML models in diagnosing heart disease, indicating its effectiveness in healthcare prediction tasks.

The study aims to achieve a twofold objective that seeks to advance the boundaries of technology within environmental science, concurrently yielding a profound impact on conservation efforts and pharmaceutical research. Initially, it focuses on the rigorous evaluation and validation of ML models, i.e., GBM, RF, and SVM, utilizing precise latitude and longitude data to ensure unparalleled locational accuracy. Subsequently, it assesses the computational efficiency of these algorithms to determine the most resource-efficient approach.

Beyond the technical accomplishments, this research contributes to ecological conservation, biodiversity protection, and pharmaceutical exploration by facilitating accurate plant geolocation. Such contributions sequentially support advanced conservation strategies, sustainable harvesting practices, and the investigation of plants' medicinal properties. This dual-focused objective highlights the study's dedication to technological innovation while underscoring its significant implications for environmental conservation and health sciences.

The structure of this study unfolds as follows: the methodology, including dataset preparation, data analysis techniques, and the particular ML models employed is detailed in Section 2; Section 3 presents the empirical results and evaluates the performance of these models; Section 4 concludes with a summary of key findings, their implications, and suggestions for future research.

2. Methodology

This study follows the research methodology illustrated in Fig. 1, starting from understanding the study's requirements to deploying efficient and accurate ML models for the precision geolocation of medicinal plants and adopting the cross-industry standard process for data mining (CRISP-DM) framework. This practical and adaptable framework [26] boosts COVID-19 diagnosis predictions [27], acts as a foundational element in ML and data science [28], and provides insights into geolocation and medicinal plant research by evaluating algorithms.

Phase 1: Business understanding: This phase explores the importance of accurately locating medicinal plants for conservation, healthcare, and sustainable development. It aims to improve geolocation with GBM, RF, and SVM algorithms, addressing the ecological research challenges and the limitations of traditional methods. The goal is to enhance conservation strategies and pharmaceutical research, expecting to advance biodiversity protection and drug discovery through precise geolocation, highlighting the study's potential to effectively bridge data science, ecology, and pharmaceuticals.

Phase 2: Data understanding: The work begins with the collection and familiarization of the geolocation data of medicinal plants. Principal component analysis (PCA) is employed in exploratory data analysis to minimize dimensionality, concurrently optimizing the dataset for ML. This step is an especially crucial factor in identifying key variables and assessing data quality, ensuring builders of subsequent phases construct them on a solid understanding of the dataset's characteristics.

Phase 3: Data preparation: This phase focused on finalizing the dataset's structure, engaging in thorough data pre-processing, including data visualization, cleaning, handling data gaps, and managing outliers. Feature engineering improves the prediction ability of models by generating new features or modifying current ones. The dataset was divided into sets for training and testing to ensure a thorough training of the model approach and enable practical performance evaluation to present new data. This preparation phase seamlessly connected data understanding with the modeling phase, establishing a robust foundation for the precise geolocation of medicinal plants through advanced ML techniques.

Phase 4: Modeling: This phase utilizes 10-fold cross-validation to assess the effectiveness of chosen ML algorithms. This approach ensures the robustness of the evaluation and no overfitting models. Furthermore, hyperparameter tuning was applied, optimizing each model's performance.

Phase 5: Evaluation: This phase assesses each model's accuracy, precision, recall, and F1 score alongside computational time efficiency metrics such as training and prediction time. This comprehensive evaluation yields a thorough comparison of the models, identifying which algorithm offers the best balance of predictive accuracy and efficiency. The interpretation of results is critical at this stage, extending insights into the models' performance and practical implications herein.

Phase 6: Deployment: Emphasizing the interpretation of results and communication to stakeholders to ensure the accessibility and feasibility of research outcomes.

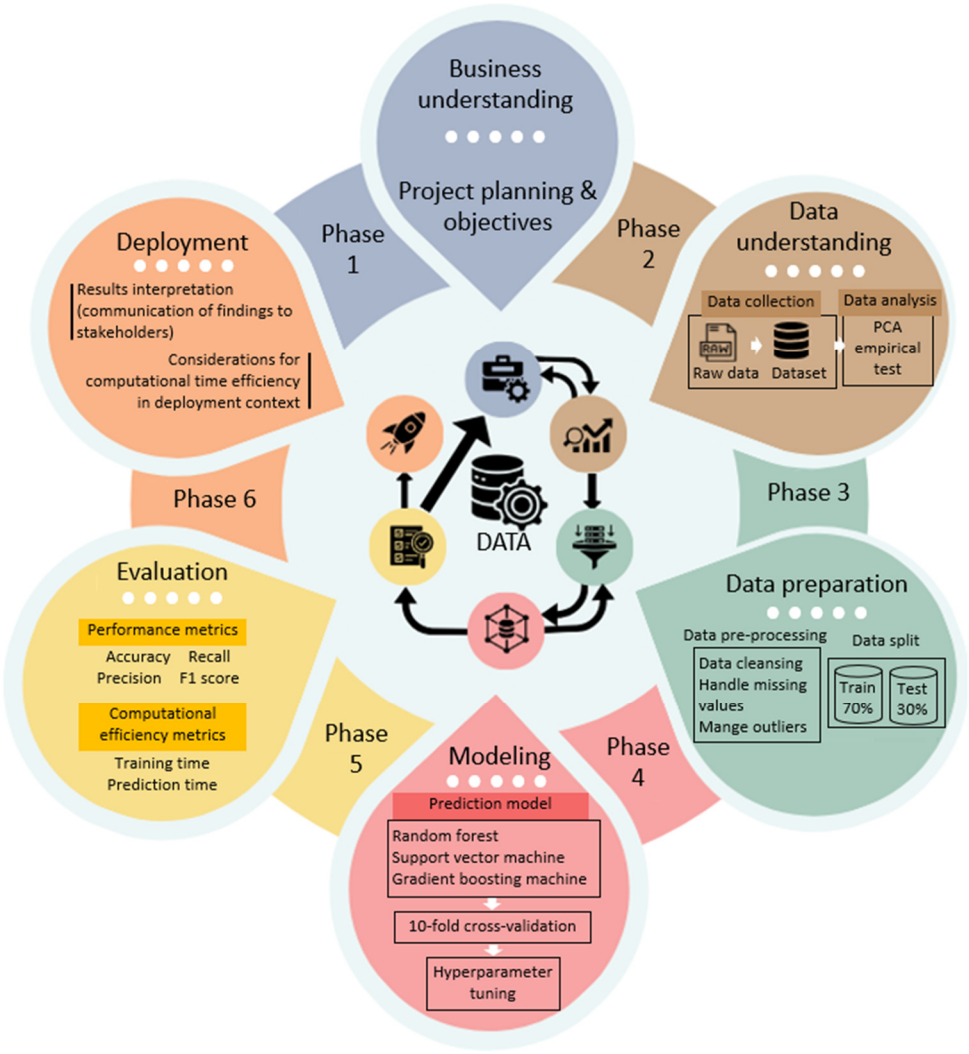


Fig. 1 Research methodology

2.1. Dataset and study area

The dataset, comprising 13 features detailed in Table 1, was curated through interviews, focus groups, online research, and fieldwork, employing diverse sampling methods for comprehensive representation. Table 2 showcases descriptive statistics from 2,212 observations on critical environmental factors crucial for evaluating ML algorithms in medicinal plant geolocation. Additionally, Fig. 2 visually represents the sample data from the dataset, proffering complementary insights.

Table 1 Dataset's feature descriptions

Feature	Description	Source
Longitude	East-west position of a plant	Fieldwork
Latitude	North-south position of a plant	Fieldwork
Temperature	Average temperature relevant to plant growth	Online
Precipitation	Levels influencing plant distribution and health	Online
Soil pH	Soil acidity or alkalinity essential for plant growth	Online
Elevation	The plant's growth environment is indicated by height above sea level	Online
MedPlantName	Names for categorizing medicinal plant species	Fieldwork
Street	The specific street number of the medicinal plant's location.	Fieldwork
Barangay	Local district or division of the plant's location.	Fieldwork
Municipality	The town or city jurisdiction of the plant's location.	Fieldwork
Province	Larger administrative division of the plant's location	Fieldwork
DateCollected	Recording date of geolocation and environmental data	Fieldwork
Present	Binary indicator of the plant's presence or absence	Fieldwork

Table 2 Medicinal plants' dataset descriptive statistics

Descriptive statistics	Latitude	Longitude	Elevation	Precipitation	Soil pH	Temperature	isPresent
Valid	2212	2212	2212	2212	2212	2212	2212
Missing	0	0	0	0	0	0	0
Mode	13.594*	124.205*	38.000*	1.311*	55.000*	28.996*	1.000*
Median	13.589	124.207	38.000	1.311	55.000	28.996	1.000
Mean	13.589	124.208	37.684	1.311	54.921	28.668	0.965
Std. deviation	0.009	0.006	13.852	0.002	0.273	0.570	0.183
Minimum	13.584	124.203	10.000	1.311	53.000	24.436	0.000
Maximum	13.871	124.230	484.000	1.380	55.000	28.996	1.000

*The mode is computed assuming that variables are discreet.

DateCollected	MedPlantName	Street	Barangay	Municipality	Province	latitude	longitude	elevation	precipitation	soilPH	temperature	isPresent
11.30.2023	Talong	None	Calatagan Proper	Virac	Catanduanes	13.5900096	124.2048396	42	1.310848713	55	28.99588785	1
11.30.2023	Akapulko	None	Calatagan Proper	Virac	Catanduanes	13.5869605	124.2044649	43	1.310848713	55	28.99588785	1
11.30.2023	Akapulko	None	Calatagan Proper	Virac	Catanduanes	13.5870859	124.2044694	43	1.310848713	55	28.99588785	1
11.30.2023	Malunggay	None	Calatagan Proper	Virac	Catanduanes	13.5900012	124.2046983	43	1.310848713	55	28.99588785	1
11.30.2023	Gumamela	Sto. Cristo	Calatagan Tibang	Virac	Catanduanes	13.5908109	124.2048346	39	1.310848713	55	28.99588785	0
11.30.2023	Bayabas	Sto. Cristo	Calatagan Tibang	Virac	Catanduanes	13.5886784	124.2046254	40	1.310848713	55	28.99588785	0
11.30.2023	Ipil-ipil	Sto. Cristo	Calatagan Tibang	Virac	Catanduanes	13.5877579	124.2045674	41	1.310848713	55	28.99588785	0
11.30.2023	Is-is	Sto. Cristo	Calatagan Tibang	Virac	Catanduanes	13.5889327	124.2045193	42	1.310848713	55	28.99588785	0
11.30.2023	Gumamela	Sto. Cristo	Calatagan Tibang	Virac	Catanduanes	13.5900552	124.2046896	43	1.310848713	55	28.99588785	0
11.30.2023	Malunggay	Sto. Cristo	Calatagan Tibang	Virac	Catanduanes	13.5892313	124.2047074	43	1.310848713	55	28.99588785	0

Fig. 2 Dataset's sample data

The study, underscoring the cruciality of the dataset in examining environmental impacts and refining the accuracy of ML algorithms, is centered on the locations of medicinal plants in three barangays within Virac, Catanduanes, Philippines—i.e., Calatagan Proper, Calatagan Tibang, and Sogod-Tibgao BLISS, as depicted in Fig. 3. Integrating geospatial, environmental, and local knowledge domains amplifies the dataset's utility, laying a sturdy groundwork for leveraging ML techniques to accomplish the research objectives effectively.

Fig. 3 Geolocations of medicinal plants in three barangays of Virac, Catanduanes

2.2. *Data analysis*

Before model development, conducting a comprehensive dataset analysis is crucial. PCA was an essential technique for reducing the dataset's complexity. As depicted in Fig. 4, PCA condensed the vast variations within the data into new, orthogonal variables, thereby revealing significant patterns. Specifically, the first principal component (PC1) accounted for 50.12% of the variance, highlighting the importance of latitude, elevation, and precipitation. These variables indicate an environmental gradient crucial for the distribution and diversity of medicinal plants, influencing their growth and characteristics. The second principal component (PC2), contributing an additional 41.17% to the variance, emphasized the roles of longitude, soil pH, and elevation, further illustrating the complex relationship between geographical and environmental factors in determining plant characteristics. Together, PC1 and PC2 encapsulated 91.29% of the total variance, effectively reducing the dimensionality of the dataset while retaining the essence of the information. This reduction facilitates easier visualization, comprehension, and application in subsequent analyses or decision-making processes related to medicinal plants.

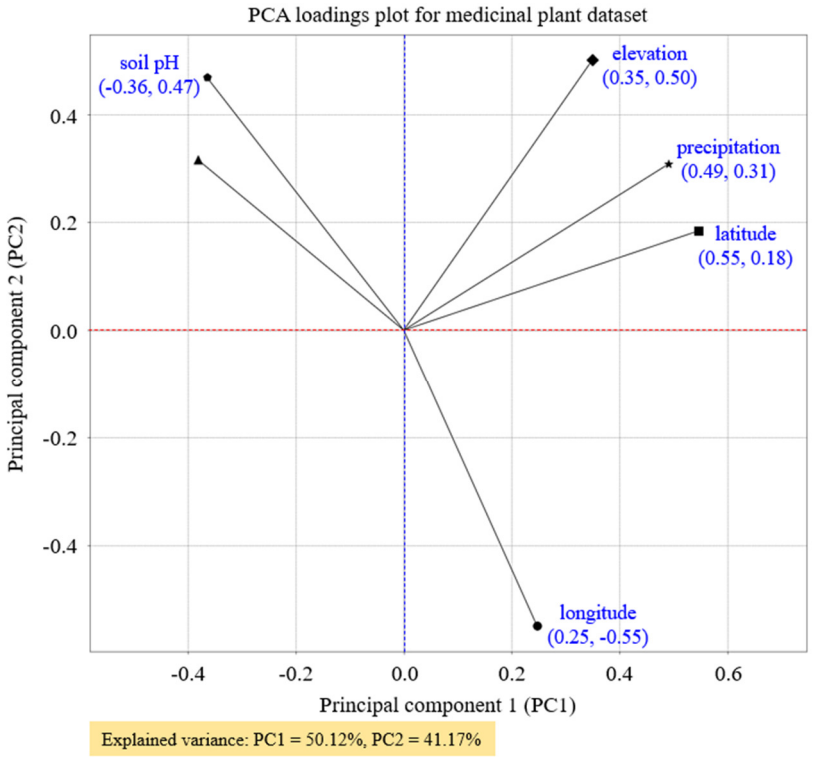


Fig. 4 Principal component analysis (PCA) results

The PCA loading plot, as detailed, underscores the significance of variables like longitude, latitude, temperature, elevation, precipitation, and soil pH in defining the traits of medicinal plants. By streamlining the dataset through this efficient dimensionality reduction, the computational efficiency of the machine learning models targeting medicinal plant geolocation was significantly enhanced. This strategic leverage of PCA insights during the model development phase enabled thorough empirical testing, assessing the influence of the identified principal components on the algorithms' predictive accuracy and computational performance. This methodology not only affirmed PCA's critical role in optimizing machine learning workflows but also highlighted its importance in studies that demand a nuanced comprehension of complex, multidimensional data for the effective conservation, analysis, and sustainable utilization of medicinal plant resources.

2.3. Data pre-processing

For ML research, it's indispensable to conduct data preprocessing, addressing discrepancies in coordinate formats and plant species identification for geographical and botanical precision. The study employs data cleansing outlier management alongside a 70% training and 30% testing split [29] for practical model fitting or training and evaluation. Through feature engineering, environmental variables are transformed into crucial predictors, emphasizing thorough preparation to improve the precision of ML and its impact on predictive analysis.

2.4. Machine learning techniques model

It is necessary to select ML algorithms like GBM, RF, and SVM for precisely geolocating medicinal plants, driven by their substantiated effectiveness and suitability for the dataset's specific characteristics. Researchers opt for these algorithms considering their capacity to accurately predict plant locations based on curated attributes, highlighting the importance of algorithm selection in enhancing research efficiency.

2.4.1. Gradient boosting machine

Deploying GBM enables researchers to iteratively enhance predictions by assembling weak models into more accurate composite models drives its selection. It excels in refining predictions for complex patterns in geospatial and environmental data related to medicinal plant habitats. Therefore, the strength is evident in intricate and nonlinear relationships between factors.

Nonetheless, the sequential and iterative nature affixed to this method requires significant computational resources, which is decisive in the efficiency objectives. Still, the potential trade-off in computation time might be warranted if GBM outpaces the accuracy of other models. Consequently, GBM emerges as a formidable option for accurately determining the geolocation of medicinal plants, promising to enhance the precision of plant location efforts and contribute positively to ecological conservation and pharmaceutical research.

2.4.2. Random forest

Due to its durability and forecast accuracy, the RF algorithm is appropriate for medicinal plant location research. It outperforms binary target variables such as medicinal plant detection and delivers accurate, precise recall and F1 score outcomes. Overfitting is prevented by RF's design, which integrates predictions from many decision trees trained on various data subsets. This feature enables the model to function with versatility and reliability.

The parallel processing capabilities of RF enhance computing performance, maximizing efficiency and model correctness. It manages high-dimensional data, analyzing complex environmental variables. It reduces model bias, ensuring reliable predictions for medicinal plant geolocation. RF can handle missing data and feature importance analysis, guiding conservation efforts and research by identifying essential plant localization characteristics.

Integrating RF into this study achieves accurate plant geolocation prediction aims and utilizes an approach that effectively manages computational requirements. RF is a highly reliable and beneficial method for advancing ecological conservation and pharmaceutical studies research, which is instrumental in accurately identifying the habitats of medicinal plants.

2.4.3. Support vector machine

SVM is crucial for medicinal plant geolocation due to its exceptional classification and versatility in processing both linear and non-linear data. SVM effectively finds the best hyperplane that maximizes the margin between classes, which helps identify medicinal plants in sundry geographical and environmental circumstances. SVM gains robustness and can classify in high-dimensional spaces through kernel functions, like the Radial Basis Function (RBF), polynomial, and sigmoid kernels. Adaptability is needed to interpret complicated geographical and environmental data patterns and ensure medicinal plant geolocation accuracy and precision. Accurate geolocation significantly underpins the success of ecological conservation efforts and pharmaceutical research in this study, which is a fact that cannot be overstated. By leveraging SVM's capabilities, the study overcomes the challenges posed by the curse of dimensionality and the need for precise categorization amidst diverse environmental factors. Essentially, SVM's ability to analyze intricate environmental patterns and precisely identify medicinal plants using geographical coordinates aligns with the study objectives, providing an advanced method to enhance ecological conservation and pharmaceutical research by utilizing enhanced geolocation techniques.

2.5. Performance and efficiency measures

The evaluation criteria for the ML algorithms GBM, RF, and SVM consist of two primary assessments: accuracy, which is crucial for making solid predictions about the locations of medicinal plants, and computational efficiency, which focuses on the training and prediction times of the algorithms. The focus lies on the imperative need for precise identification of potential habitats of medicinal plants and the significance of maximizing computing resources to guarantee the practical implementation of these models. These evaluation criteria form a critical assessment that balances geolocation accuracy with operational efficiency, thereby highlighting the approaches' usefulness and applicability for real-world scenarios.

2.5.1. Performance measure

Accuracy is crucial for evaluating the performance of GBM, RF, and SVM to predict the dataset's medicinal geographical locations. A high degree of accuracy indicates that the model is dependable in its predictions, perceived to be valuable for researchers and conservationists to pinpoint locations where medicinal plants are probable. In this study, a model with high accuracy would effectively differentiate between locations where medicinal plants exist and locations where they do not.

$$Accuracy\ (\%) = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \tag{1}$$

where *TP* is the true positive, *TN* is the true negative, *FP* is the false positive, and *FN* is the false negative.

Precision is essential for ensuring the accuracy of prediction concerning the existence of medicinal plants in a particular location. Precision is vital in conservation efforts with limited resources to focus efforts and resources on locations with medicinal plants, optimizing conservation actions and research treatments.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall high recall is crucial for models to identify the maximum number of real medicinal plant sites accurately. A model's high recall indicates its effectiveness in reducing false negatives, meaning it scarcely fails to identify regions where medicinal plants are found. This effectiveness is essential for thorough conservation planning and study to prevent any possible areas of interest from being missed.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The *F1 score* is a numerical representation that harmonizes precision and recall, serving as a critical gauge in situations where the consequences of both false positives and false negatives are significant. When dealing with the geolocation of medicinal plants, the F1 score proffers assistance in the selection of a model equalizing between high recall (not missing probable plant locations) and high precision (not erroneously recognizing irrelevant places as holding medicinal plants).

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

These metrics are particularly relevant herein in assessing the models' precision in geolocating medicinal plants. In this task, identifying true locations (recall) and evading false identifications (precision) are equally crucial for conservation and pharmaceutical research.

2.5.2. Efficiency measure

Computational efficiency in ML is adopted to evaluate time and resource demands for training and making predictions. It is assessed through time complexity, dataset size, and resource utilization, which measures the amount of computational resources consumed during operation. These metrics provide a comprehensive view of an algorithm's efficiency. Training and prediction times are crucial metrics for evaluating the computational efficiency of ML algorithms like GBM, RF, and SVM. Training time is when a model learns from a training dataset, enabling it to adjust parameters and improve accuracy. Shorter training times indicate faster development and deployment. Prediction time measures the model's responsiveness to new data, especially in real-time decision-making applications. Training time is a one-time cost during the learning phase, while prediction time is ongoing, affecting the model's operational efficiency. Optimizing these metrics without compromising accuracy is essential for developing accurate tools for medicinal plant geolocation.

2.6. 10-fold cross-validation

Implementing 10-fold cross-validation is to elevate ML models like GBM, RF, and SVM validation processes. Partitioning the dataset into ten equally sized segments and cyclically using these partitions for training and testing ensures an exhaustive evaluation of model performance across varied geographical and climatic conditions. Such a method mitigates bias and variance to ensure model reliability and stability. Consequently, this meticulous validation approach enhances the study's methodological rigor. It contributes valuable insights into ecological conservation and pharmaceutical research, substantiating the scientific robustness and practicability of the findings in medicinal plant conservation.

2.7. Hyperparameter tuning

After employing 10-fold cross-validation, adjustments were made to parameters including the number of trees for GBM and RF, GBM's tree depth, RF's feature considerations for splits, and SVM's kernel type and regularization parameter, all aimed to optimize enhanced predictive accuracy and overfitting prevention. This approach guaranteed model robustness and accurate prediction of medicinal plant locations across varied geospatial contexts. The results of these hyperparameter tuning efforts and the corresponding mean scores from the 10-fold cross-validation are concisely presented in Table 3, illustrating the systematic refinement and its impact on the models' performance. This study utilized Python libraries to enhance workflow efficiency, from data preprocessing to performance metrics analysis. The library called "scikit-learn" facilitated ML tasks, "pandas" and "numpy" managed data manipulation, and "matplotlib" and "seaborn" supported visualization. These tools enabled practical hyperparameter tuning and 10-fold cross-validation to improve result analysis and interpretation.

Table 3 Results of hyperparameter tuning and the mean scores for 10-fold cross-validation

Model	Best parameters		10-fold cross-validation (mean)			
	Hyperparameter	Values	Accuracy	Precision	Recall	F1 score
SVM	C	0.1	0.9652	0.9652	1.0000	0.9823
	Gamma	Scale				
	Kernel	RBF				
GBM	learning_rate	0.01	0.9557	0.9662	0.9887	0.9771
	max_depth	3				
	n_estimators	100				
RF	max_depth	10	0.9566	0.9662	0.9897	0.9776
	n_estimators	500				

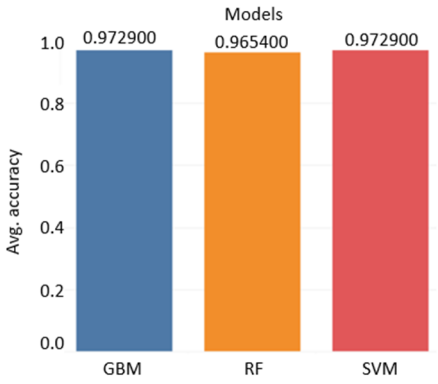
3. Results and Discussion

This section analyzes ML models, examining the accuracy and computational efficiency in determining medicinal plant locations. Four performance metrics and two efficiency measures, as shown in Table 4, were used to assess each model's capabilities. The primary goal of this research is to investigate the efficacy of GBM, RF, and SVM algorithms concerning the capability to geolocate medicinal plants using geospatial data, focusing on their precision and processing speed.

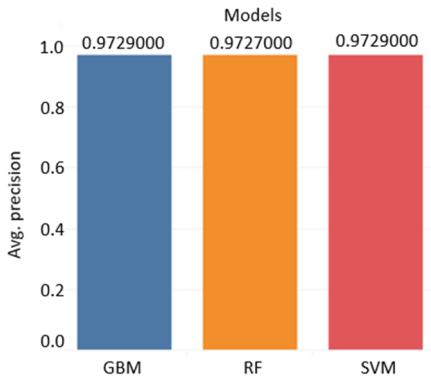
Table 4 Summary of results in final model training and evaluation

Model	Performance measure				Computational efficiency (in seconds)	
	Accuracy	Precision	Recall	F1 score	Training time	Prediction time
SVM	97.29%	97.29%	100.00%	98.63%	0.008	0.0040
GBM	97.29%	97.29%	100.00%	98.63%	0.203	0.0012
RF	96.54%	97.27%	99.23%	98.24%	1.597	0.0601

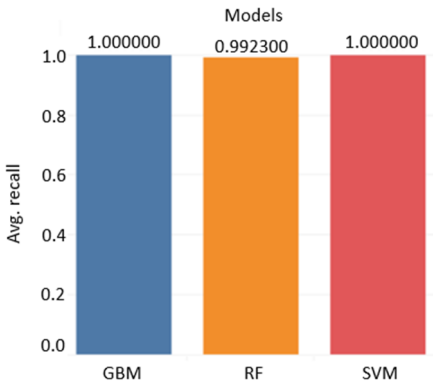
Fig. 5 and Fig. 6 present a graphical comparison of the performance and computational efficiency metrics across the models. This visualization highlights the impressive performance and computational efficiency of the algorithm, exhibiting proficiency in accurately determining the locations of medicinal plants.



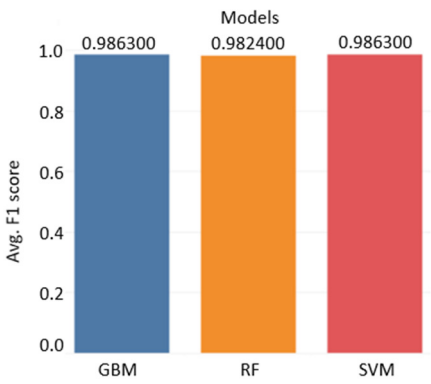
(a) Comparison of accuracy among models



(b) Comparison of precision among models

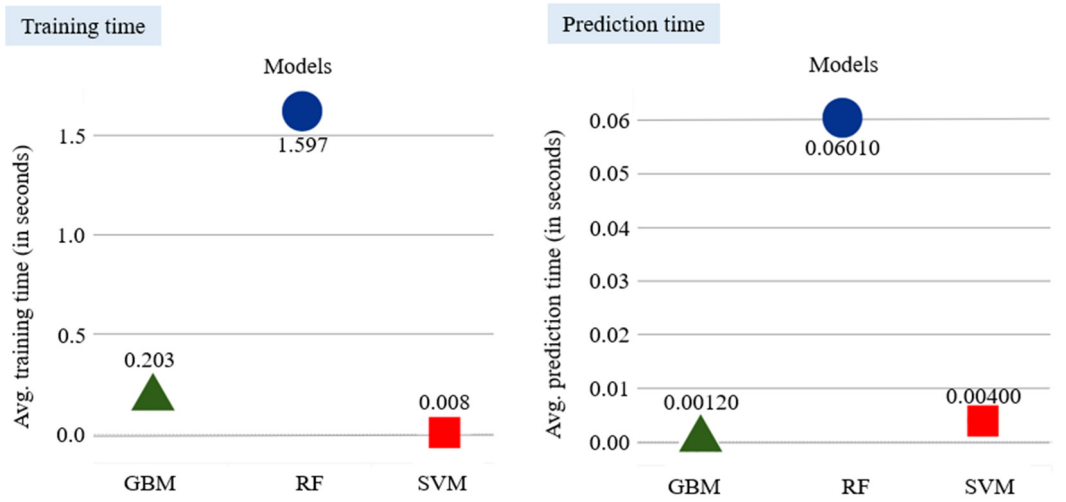


(c) Comparison of recall among models



(d) Comparison of F1 score among models

Fig. 5 Comparison of performance measures among models



(a) Comparison of training time among models (b) Comparison of prediction time among models

Fig. 6 Model comparison of computational efficiency

3.1. Comparative evaluation of the performance of models

Fig. 5 demonstrates the similarity reflecting on the performance levels between SVM and GBM models, achieving 97.29% accuracy and precision, a 100.00% recall rate, and a 98.63% F1 score. These metrics underscore the exceptional proficiency of both the SVM and GBM algorithms in accurately detecting the presence of medicinal plants, ensuring a high success rate in correct identifications while minimizing false positives. On the other hand, while slightly trailing, the RF model manifests robust performance, achieving 96.54% in accuracy, 97.27% in precision, 99.23% in recall, and 98.24% in the F1 score. Despite a subtle decrease in performance metrics compared to SVM and GBM, the RF model maintains its status with convincing competitiveness, exhibiting considerable efficiency and reliability in medicinal plant geolocation.

SVM and GBM algorithms evinced superior, closely matched performance suitable for ecological conservation and pharmaceutical research, highlighted by perfect recall rates critical for comprehensive geographical mapping. The dataset's traits, dimensionality reduction, hyperparameter choices, or the binary classification nature could result in fluctuation in the performance. These findings indicate that exploring different configurations might reveal nuanced performance differences, guiding future research to enhance medicinal plant geolocation precision.

Despite displaying slightly lower performance metrics, the RF model is a robust contender, attesting to considerable accuracy and precision in identifying medicinal plant locations. Compared to SVM and GBM, its marginally lower recall rate insignificantly detracts from its utility, especially in scenarios where model interpretability and the capability to navigate complex data structures are paramount. Highlighting RF's utility underscores the necessity of selecting the most fitting ML algorithm to satisfy the unique demands of a project, thus ensuring the advancement of medicinal plant geolocation in both efficiency and accuracy.

3.2. Comparative evaluation of computational efficiency of models

This study assessed the computational efficiency by analyzing the training and prediction times. As detailed in Fig. 6, the SVM showed exceptional efficiency, with a training time of 0.008 seconds and a prediction time of 0.004 seconds, which is perceived as the quickest model among those evaluated. The GBM demonstrated exceptional efficiency during training, completing in 0.203 seconds and having an even more remarkable prediction time of 0.0012 seconds, the fastest of the three methods. The RF model had the most significant training time of 1.597 seconds and a prediction time of 0.060 seconds, signifying lesser performance than other models but still viable for many applications.

This study's evaluation of SVM, GBM, and RF algorithms for medicinal plant geolocation uncovers unique performance traits essential for tailored application needs. With its rapid training and prediction capabilities, the SVM algorithm is particularly beneficial for applications demanding immediate responsiveness, such as mobile or real-time geolocation tasks. This promptness in processing makes SVM an invaluable asset in environments where speed is paramount, directly contributing to the timely and efficient geolocation of medicinal plants. On the other hand, GBM demonstrates a subtly slower training pace but excels in prediction speed, positioning it as the preferred choice for scenarios requiring continuous model updates and instantaneous decision-making. This unique balance between training duration and predictive velocity underscores GBM's suitability for dynamic applications, where quick adjustments based on new data are essential.

Conversely, the RF model's extended training period indicates its applicability when models can be pre-trained offline. The benefits of model interpretability and nuanced feature interaction comprehension outweigh the necessity for swift computation. Such characteristics suggest RF's potential in comprehensive studies or applications where the depth of analysis and accuracy precedes immediate results. The comparative evaluation underscores the critical role of aligning algorithm selection with operational demands, including speed, accuracy, and computational resource availability considerations. Highlighting the specific advantages of each algorithm in the context of medicinal plant geolocation attests to the importance of selecting the appropriate ML approach to tackle the unique challenges of precision geolocation. It ensures that the selection process accounts for both computational efficiency and practical implications for real-world applications.

4. Conclusions

The study aimed to assess the accuracy and computational performance of three ML techniques—GBM, RF, and SVM—for the geolocation of medicinal plants using spatial data. It adhered to the CRISP-DM framework, ensuring a structured evaluation from project understanding to model deployment. The SVM and GBM models showcased superior performance in identifying medicinal plant locations, with the SVM model achieving notable efficiency in training and prediction times, which is suitable for real-time applications. The GBM model was highlighted for its quick prediction capabilities, while the RF model was recommended for scenarios demanding high interpretability and complex feature interaction management. Performance metrics for SVM and GBM included a 97.29% accuracy rate, 100% recall, and a 98.63% F1 score, indicating their exceptional capability in precise geolocation tasks. The study underscored the SVM model's computational efficiency, making it an optimal choice for applications that necessitate quick and reliable predictions.

The findings have significant implications for environmental conservation and healthcare, aiding in the accurate geolocation of medicinal plants. This supports targeted conservation efforts and sustainable harvesting, which is crucial for preserving biodiversity and continuing traditional medicinal knowledge.

The study advocates for further integrating more advanced ML models or exploring deep learning techniques to enhance geolocation accuracy and efficiency. It posits that such advancements could revolutionize precision geolocation, contributing notably to ecological conservation and the pharmaceutical industry by marrying technological innovation with practical applications.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] C. S. Cordero, U. Meve, and G. J. D. Alejandro, "Ethnobotanical Documentation of Medicinal Plants Used by the Indigenous Panay Bukidnon in Lambunao, Iloilo, Philippines," Frontiers in Pharmacology, vol. 12, article no. 790567, January 2022.

[2] O. Nuneza, B. Rodriguez, and J. G. Nasiad, "Ethnobotanical Survey of Medicinal Plants Used by the Mamanwa Tribe of Surigao Del Norte and Agusan Del Norte, Mindanao, Philippines," Biodiversitas Journal of Biological Diversity, vol. 22, no. 6, pp. 3284-3296, June 2021.

[3] J. M. Lopez and J. M. Tram, "Falling Behind and Forgotten: The Impact of Acculturation and Spirituality on the Mental Health Help-Seeking Behavior of Filipinos in the USA," Asian American Journal of Psychology, vol. 14, no. 2, pp. 218-230, 2023.

[4] H. S. Faizy, G. Y. Haji, S. M. Saeed, T. S. Mala, M. S. Ibrahim, M. A. Khider, et al., "Geographical Study of Medicinal Plants Using GIS and GPS Tools in Some Villages, Barzan Sub-District, Mergasor Districts Iraqi Kurdistan Region," IOP Conference Series: Earth and Environmental Science, vol. 1252, article no. 012174, December 2023.

[5] L. R. Halpin, J. D. Ross, R. Ramos, R. Mott, N. Carlile, N. Golding, et al., "Double-Tagging Scores of Seabirds Reveals that Light-Level Geolocator Accuracy is Limited by Species Idiosyncrasies and Equatorial Solar Profiles," Methods in Ecology and Evolution, vol. 12, no. 11, pp. 2243-2255, November 2021.

[6] P. A. Singh, A. Sood, and A. Baldi, "An Agro-Ecological Zoning Model Highlighting Potential Growing Areas for Medicinal Plants in Punjab," Indian Journal of Pharmaceutical Education and Research, vol. 55, no. 2s, pp. s492-s500, April-June 2021.

[7] W. A. R. W. M. Isa, I. M. Amin, and N. Saubiran, "Mobile Application on Malay Medicinal Plants Based on Information Crowdsourcing," Alinteri Journal of Agriculture Sciences, vol. 36, no. 2, pp. 208-229, 2021.

[8] D. Sugiarto, J. Siswantoro, M. F. Naufal, and B. Idrus, "Mobile Application for Medicinal Plants Recognition from Leaf Image Using Convolutional Neural Network," Indonesian Journal of Information Systems, vol. 5, no. 2, pp. 43-56, February 2023.

[9] S. Puttinaovarat and P. Horkaew, "A Geospatial Database Management System for the Collection of Medicinal Plants," Geospatial Health, vol. 16, no. 2, article no. 998, October 2021.

[10] R. Permana, E. T. Tosida, and M. I. Suriansyah, "Development of Augmented Reality Portal for Medicininal Plants Introduction," International Journal of Global Operations Research, vol. 3, no. 2, pp. 52-63, 2022.

[11] M. O. Faruque, G. Feng, M. N. A. Khan, J. W. Barlow, U. R. Ankhi, S. Hu, et al., "Qualitative and Quantitative Ethnobotanical Study of the Pangkhua Community in Bilaichari Upazilla, Rangamati District, Bangladesh," Journal of Ethnobiology and Ethnomedicine, vol. 15, article no. 8, 2019.

[12] P. Boycheva and D. Ivanov, "Comparative Ethnobotanical Analysis of the Used Medicinal Plants in the Region of the Northern Black Sea Coast (Bulgaria)," Acta Scientifica Naturalis, vol. 8, no. 2, pp. 44-54, July 2021.

[13] K. S. M. Anbananthen, S. Subbiah, D. Chelliah, P. Sivakumar, V. Somasundaram, K. H. Velshankar, et al., "An Intelligent Decision Support System for Crop Yield Prediction Using Hybrid Machine Learning Algorithms," F1000Research, vol. 10, article no. 1143, November 2021.

[14] Y. M. Chen, Y. Kao, C. C. Hsu, C. J. Chen, Y. Ma, Y. S. Shen, et al., "Real-Time Interactive Artificial Intelligence of Things-Based Prediction for Adverse Outcomes in Adult Patients with Pneumonia in the Emergency Department," Academic Emergency Medicine, vol. 28, no. 11, pp. 1277-1285, November 2021.

[15] P. Theerthagiri and J. Vidya, "Cardiovascular Disease Prediction Using Recursive Feature Elimination and Gradient Boosting Classification Techniques," Expert Systems, vol. 39, no. 9, article no. e13064, November 2022.

[16] C. A. ul Hassan, J. Iqbal, S. Hussain, H. AlSalman, M. A. A. Mosleh, and S. Sajid Ullah, "A Computational Intelligence Approach for Predicting Medical Insurance Cost," Mathematical Problems in Engineering, vol. 2021, article no. 1162553, 2021.

[17] S. Mohapatra and N. Chaudhary, "Statistical Analysis and Evaluation of Feature Selection Techniques and Implementing Machine Learning Algorithms to Predict the Crop Yield Using Accuracy Metrics," Engineered Science, vol. 21, article no. 787, February 2023.

[18] M. Waleed, T. W. Um, T. Kamal, and S. M. Usman, "Classification of Agriculture Farm Machinery Using Machine Learning and Internet of Things," Symmetry, vol. 13, no. 3, article no. 403, March 2021.

[19] H. H. Wang, C. C. Huang, P. C. Talley, and K. M. Kuo, "Using Healthcare Resources Wisely: A Predictive Support System Regarding the Severity of Patient Falls," Journal of Healthcare Engineering, vol. 2022, article no. 3100618, 2022.

[20] P. Michailidis, A. Dimitriadou, T. Papadimitriou, and P. Gogas, "Forecasting Hospital Readmissions with Machine Learning," Healthcare, vol. 10, no. 6, article no. 981, June 2022.

[21] H. Oh, "A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model," IEEE Access, vol. 9, pp. 144121-144128, 2021.

[22] D. J. Pangarkar, R. Sharma, A. Sharma, and M. Sharma, "Assessment of the Different Machine Learning Models for Prediction of Cluster Bean (Cyamopsis Tetragonoloba L. Taub.) Yield," Advances in Research, vol. 21, no. 9, pp. 98-105, 2020.

[23] C. Sawangwong, K. Puangsuwan, N. Boonnam, S. Kajornkasirat, and W. Srisang, "Classification Technique for Real-Time Emotion Detection Using Machine Learning Models," IAES International Journal of Artificial Intelligence (IJ-AI), vol. 11, no. 4, pp. 1478-1486, December 2022.

[24] L. Shi, Y. Qin, J. Zhang, Y. Wang, H. Qiao, and H. Si, "Multi-Class Classification of Agricultural Data Based on Random Forest and Feature Selection," Journal of Information Technology Research, vol. 15, no. 1, pp. 1-17, 2022.

[25] T. Suresh, T. A. Assegie, S. Rajkumar, and N. K. Kumar, "A Hybrid Approach to Medical Decision-Making: Diagnosis of Heart Disease with Machine-Learning Model," International Journal of Electrical and Computer Engineering, vol. 12, no. 2, pp. 1831-1838, April 2022.

[26] F. Martinez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernandez-Orallo, M. Kull, N. Lachiche, et al., "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 8, pp. 3048-3061, August 2021.

[27] J. S. Saltz and I. Krasteva, "Current Approaches for Executing Big Data Science Projects-A Systematic Literature Review," PeerJ Computer Science, vol. 8, article no. e862, 2022.

[28] D. Oliveira, D. Ferreira, N. Abreu, P. Leuschner, A. Abelha, and J. Machado, "Prediction of COVID-19 Diagnosis Based on OpenEHR Artefacts," Scientific Reports, vol. 12, article no. 12549, 2022.

[29] S. Montaha, S. Azam, A. K. M. R. H. Rafid, S. Islam, P. Ghosh, and M. Jonkman, "A Shallow Deep Learning Approach to Classify Skin Cancer Using Down-Scaling Method to Minimize Time and Space Complexity," PLoS ONE, vol. 17, no. 8, article no. e0269826, 2022.