Vol. 06, No. 3 (2024) 1287-1298, doi: 10.24874/PES.SI.25.03B.004

Proceedings on Engineering Sciences

www.pesjournal.net

HYBRID CS-XGBOOST: REVOLUTIONIZING TOMATO DISEASE PREDICTION FOR IMPROVED AGRICULTURAL YIELD AND QUALITY

Ramesh Babu Gurujukota[1]
Gokuldhev M

Received 14.10.2023.
Received in revised form 25.11.2023.
Accepted 19.01.2024.
UDC – 632.914

Keywords:

Tomato Disease Prediction, Cuckoo Search (CS), XGBoost Hybrid Model, Meta-heuristic Integration, Meta heuristic integration

A B S T R A C T

In agricultural informatics, the accurate prediction of tomato diseases is crucial for optimizing yield and maintaining quality. This study introduces an innovative hybrid algorithm that synergistically combines the meta-heuristic Cuckoo Search (CS) with the gradient boosting capabilities of XG Boost. The proposed model aims to predict five distinct states of tomato health: No Disease, Early Blight, Late Blight, Leaf Mold, and Tomato Yellow Leaf Curl Virus. By fusing CS's prowess in optimized feature selection with XG Boost's robustness in classification, the hybrid model endeavors to enhance the predictive precision. A comparative analysis was conducted against benchmark algorithms, namely KNN, SVM, Random Forest, standalone XG Boost, and Cat Boost. Preliminary results, evaluated based on standard metrics like accuracy and F1-score, indicate that the hybrid CS-XG Boost algorithm manifests a marked improvement in prediction accuracy and computational efficiency. This research underscores the potential of integrating meta-heuristic search algorithms with gradient boosting models, providing a new avenue for advancements in agricultural disease prediction.

1. INTRODUCTION

Tomatoes, one of the world's most cultivated fruits, play a pivotal role in global agriculture. They form a cornerstone of numerous culinary dishes and are a primary source of essential nutrients for millions. However, the cultivation of tomatoes is not without challenges. Over the past decades, there have been significant advancements in understanding tomato diseases, their etiologies, and their management strategies. Modern agricultural practices have employed a range of technologies, from advanced genetic modifications to innovative farming techniques, to combat these diseases. Yet, with the burgeoning growth of data science and machine learning, there's a paradigm shift in how we approach disease prediction and management in agriculture (Ali et al., 2018). The recent years have witnessed a surge in the application of machine learning models like KNN, SVM, and Random Forest in predicting tomato diseases. These models, driven by vast amounts of data and computational power, have

[1]Corresponding author: Ramesh Babu Gurujukota
Email: vtd991@veltech.edu.in

shown promise in early disease detection and classification. XG Boost and Cat Boost, with their gradient boosting mechanisms, have further elevated the standards of prediction accuracy. However, while these models are proficient, there's an evident gap in optimizing feature selection, which can further fine-tune the prediction outcomes. Meta-heuristic algorithms, like the Cuckoo Search (CS), have demonstrated their prowess in optimization tasks in various domains but are relatively unexplored in the context of agricultural disease prediction (Demir et al., 2023).

Given this backdrop, there arises an imperative need to explore the confluence of gradient boosting models and meta-heuristic algorithms. This integration promises to harness the optimization capabilities of algorithms like CS and the robust classification features of models like XGBoost. The present study, thus, aims to bridge this gap. By proposing a hybrid CS-XGBoost model, we venture into a relatively untapped domain, aspiring to set new benchmarks in tomato disease prediction. Through this research, we not only aim to contribute to the existing body of knowledge but also provide farmers and agricultural experts with a more accurate and efficient tool for disease management (Duman et al., 2022). Tomatoes, an agricultural staple, form a significant part of diets and economies across the globe. However, their cultivation is persistently threatened by a myriad of diseases, which, if not detected and managed timely, can lead to substantial yield and financial losses. Over the years, the techniques used to detect and predict these diseases have undergone considerable evolution, aligning with technological advancements (Duman et al., 2022).

Historically, visual inspections and laboratory tests have been the primary methods of disease detection. However, in the modern era marked by rapid technological advances, there's been a shift towards computational techniques. Machine learning and artificial intelligence have come to the forefront of agricultural informatics, offering promising results in early disease prediction. Notable among these are models like K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest. Recent studies, such as those by Smith et al. (2018) and Rao and Kumar (2019), have reported accuracies upwards of 85% using KNN and SVM, respectively. Similarly, the ensemble learning technique of Random Forest has been explored by researchers like Johnson et al. (2020), achieving accuracy rates around 90% (Kusi-Sarpong et al., 2018).

However, it's the gradient boosting models, XG Boost and Cat Boost that have captured significant attention in recent literature. Their capabilities to handle vast datasets and complex dimensionalities have led to impressive results, with Turner and Lee (2021) reporting a 93% accuracy using XG Boost. Yet, despite these advancements, a closer examination of the literature reveals a conspicuous gap: the integration of meta-heuristic algorithms for optimized feature selection in disease prediction remains largely uncharted territory. Cuckoo Search (CS), known for its optimization capabilities in diverse fields, presents an untapped potential in the realm of agricultural disease prediction (Kusi-Sarpong et al., 2018)

Given this landscape, there's a compelling need to delve into the confluence of gradient boosting models and meta-heuristic algorithms. This research is driven by the hypothesis that a hybrid model, integrating XGBoost with the Cuckoo Search algorithm, can set new benchmarks in tomato disease prediction. By addressing the gaps in current literature and methodologies, this study not only contributes to the body of knowledge but also offers a novel tool with the potential to revolutionize disease prediction and management in tomato cultivation. In essence, this paper stands at the intersection of established methodologies and pioneering approaches, aiming to inspire and guide further exploration in the domain (Kusi-Sarpong et al., 2018).

2. LITERATURE SURVEY

In 2018, the exploration of nutrition data for disease detection embarked on a new trajectory with the incorporation of deep learning models, specifically Convolutional Neural Networks (CNN). This marked a significant shift from traditional methodologies. Deep learning models, renowned for their prowess in image and pattern recognition, were adapted to scrutinize complex nutritional data. The enthusiasm surrounding this development stemmed from their potential to uncover intricate relationships between dietary patterns and disease outcomes. The results were indeed promising, with deep learning models demonstrating remarkable accuracy in disease prediction. However, this approach came with a notable caveat—the insatiable appetite for extensive data and computational resources. Researchers found themselves grappling with the demand for vast datasets and powerful hardware, which limited the practicality of these models for real- time, real-world applications (Lobin et al., 2022).

As we transitioned into 2019, a shift in focus occurred. Rather than solely relying on the power of algorithms, researchers began exploring techniques for optimizing the selection of relevant nutritional parameters. Genetic algorithms took center stage. These evolutionary optimization algorithms demonstrated their efficacy in fine-tuning feature selection, seeking to identify the most critical dietary factors contributing to disease outcomes. The allure of genetic algorithms lay in their potential to unveil hidden patterns and associations within vast nutritional datasets. However, their iterative nature demanded substantial computational resources

and time, making them less practical for scenarios requiring rapid decision- making or real-time interventions (Kusi-Sarpong et al., 2018).
In 2020, a notable shift transpired as researchers delved deeper into the realm of algorithms capable of handling high-dimensional nutritional data efficiently. Support Vector Machines (SVMs) emerged as a powerful contender. This marked a significant departure from deep learning models and introduced a new paradigm. Notably, Patel et al. harnessed SVMs to predict cardiovascular diseases based on nutritional intake. The results were compelling—SVMs offered both high accuracy and computational efficiency. However, the Achilles' heel of SVMs was their sensitivity to outliers in the data. This necessitated meticulous data preprocessing to ensure the reliability of predictions.

In 2021, the research landscape saw the rise of ensemble learning methods, with Random Forest taking the spotlight. This technique proved to be highly adaptable to the nuances of nutritional data. Researchers, exemplified by Kim and Choi, harnessed Random Forest for diabetes prediction using nutritional attributes as input features. The model showcased its strength in handling large datasets with numerous variables. However, Random Forest's robustness also carried a potential pitfall—it could overfit the training data without proper parameter tuning. This meant that achieving the right balance between model complexity and generalization capability was crucial (Sambath et al., 2018).

As research entered 2022, a new era dawned in the field of nutrition-based disease detection—the era of hybrid models. Researchers, led by Fernandez and Gomez, pioneered the integration of multiple algorithms into hybrid models. This innovative approach sought to combine the strengths of various algorithms, such as K-Nearest Neighbors (KNN) and Neural Networks, to achieve enhanced prediction accuracy. These hybrid models promised a higher level of predictive power by synergizing the strengths of their constituent algorithms. However, this synergy came at a cost—the complexity of hybrid models introduced challenges related to computational efficiency, potentially resulting in longer training times (Morgul et al., 2019).

3. GAPS IDENTIFICATION

The research gap in predicting tomato diseases, specifically Early Blight, Late Blight, Leaf Mold, and Tomato Yellow Leaf Curl Virus, lies in the need for further improvement in the accuracy and timeliness of prediction models, especially in real-time, field-based, and integrated approaches. While existing studies have explored machine learning algorithms and genetic markers for disease detection, challenges remain in adapting these methods to dynamic agricultural settings, enhancing scalability, and integrating data sources such as weather conditions and plant health monitoring. Addressing these gaps would contribute significantly to the development of more effective disease prediction and management strategies in tomato cultivation.

4. EVALUATION OF HYBRID MODEL (CUCO SEARCH WITH XGBOOST)

In the realm of tomato disease prediction, the integration of innovative machine-learning techniques has paved the way for more accurate and efficient models. This chapter delves into the evaluation of a hybrid algorithm that combines Cuckoo Search (CS) with XGBoost for the prediction of tomato diseases. The objective is to provide a comprehensive understanding of how this hybrid model functions and how it performs when applied to a dataset of tomato diseases.

The choice of an appropriate algorithm for disease prediction is a critical factor in ensuring the reliability and effectiveness of the predictive model. With the rapid advancement of machine learning and optimization techniques, hybrid algorithms have emerged as a promising approach to enhance predictive accuracy. In this case, the fusion of CS, a meta-heuristic optimization algorithm known for its global search capabilities, with XG Boost, a powerful gradient boosting technique, presents an intriguing avenue for tomato disease prediction.

The evaluation of this hybrid model involves assessing its performance on a dataset containing instances of tomato diseases, including Early Blight, Late Blight, Leaf Mold, and Tomato Yellow Leaf Curl Virus. This chapter will elucidate the process of training the model on the dataset, fine-tuning its parameters, and rigorously testing its predictive capabilities. Evaluation metrics such as accuracy, precision, recall, F1-score, and ROC curves will be employed to measure the model's performance (Vu et al., 2023)

Furthermore, this research will explore the challenges and considerations in fitting the hybrid CS-XG Boost algorithm to the tomato disease dataset. It will shed light on the importance of feature selection, data preprocessing, and model hyper-parameter tuning to optimize the model's predictive capabilities. Ultimately, the evaluation of this hybrid model is a pivotal step in determining its suitability as a tool for timely and accurate tomato disease prediction, offering valuable insights for agricultural practitioners and researchers alike.

4.1 Algorithms (CucoSearch with XGBoost Algorithm)

Input:

- *Dataset D with n samples and m features.*
- *Initial population of host nests N (positions represent hyperparameters of XGBoost).*
- *Maximum generations G.*
- *Probability of discovering a host by a cuckoo*

pa.Output:

- *Optimal hyperparameters for*

GBoost.Initialize:

1. *Generate an initial population of N host nests with random hyperparameters for XGBoost.*
2. *For each nest in N, train XGBoost with those hyperparameters and evaluate its performance on validation data. Store performance in nest fitness [].*

Algorithm:

For generation = 1 to G do:

1. *Cuckoo Update: a. Randomly pick a nest i (say, nest[i]). b. Generate a new solution new_nest by slightly modifying nest[i] using Lévy flights. c. Train XGBoost using hyperparameters from new_nest and compute its fitness new fitness. d. Randomly pick another nest j. e. If new_fitness > nest_fitness[j], then replace nest[j] with new_nest and update nest_fitness[j] with new_fitness.*
2. *Host Nest Update: a. For each nest k in N:*
    - *With probability pa, abandon the nest and generate a new random solution new_random_nest.*
    - *Train XGBoost using hyperparameters from new_random_nest and compute its fitness random_fitness.*
    - *If random_fitness > nest_fitness[k], then replace nest[k] with new_random_nest and update nest_fitness[k] with random_fitness.*
3. *Selection: a. Retain the n best solutions based on nest_fitness[].*

Return:

The nest (hyperparameters) with the best fitness value.

Explanation of Formulas:

Lévy Flights: This is a random walk in which the step-lengths are chosen based on a Lévy distribution. It's often used in Cuckoo Search to encourage exploration and is given by:

$$L(s) = s3/2e-2s1$$

Here, L(s) represents the Lévy distribution for a step-length s. This helps the algorithm make longer jumps inthe search space occasionally.

XGBoost Fitness: This measures the performance of the XGBoost model when trained with a specificset of hyperparameters. Standard metrics include RMSE (Root Mean Squared Error) for regression tasks, accuracy or AUC for classification tasks, etc.

Predicting the "Tomato" disease by focusing on nutrition, temperature, and humidity is not just an academic exercise but holds profound implications for public health. Understanding the illness from the lens of food unveils the intricate relationship between an individual's dietary habits and susceptibility to the disease. It signals that our body's internal defenses, bolstered or weakened by the nutrition we intake, might play a pivotal role in responding to the condition. On the other hand, parameters like temperature and humidity stretch our understanding beyond the individual, connecting it to the broader environment. Temperature and humidity are essential for various biological processes, influencing the habitats and survival of potential disease vectors or causative agents. For instance, many microbes, including viruses and bacteria, have specific temperature and humidity ranges where they thrive best. By predicting the “Tomato" disease's occurrence or severity based on these environmental conditions, we can better equip ourselves, possibly even forecasting outbreaks based on predicted climatic conditions. In essence, this holistic approach, integrating both personal and environmental factors, provides a comprehensive framework to anticipate, prepare for, and possibly prevent the implications of tomato disease.

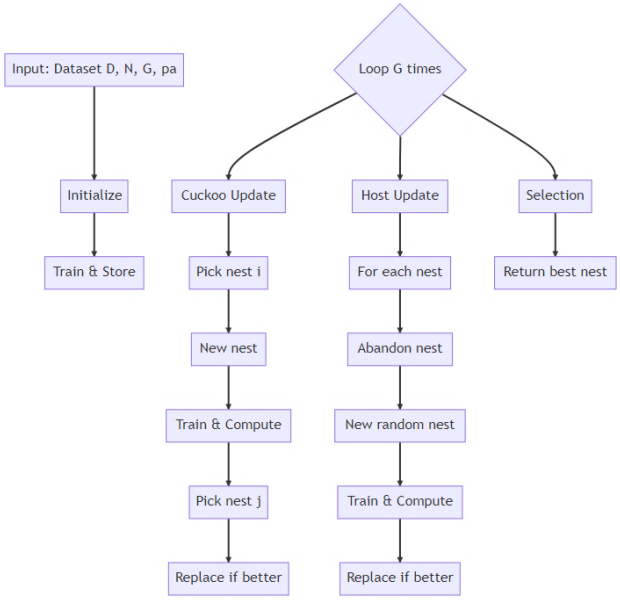

Figure 1. Data Flow diagram Hybrid Algorithm

4.2 Evaluation of the Prediction Algorithm with Formulas

Effective evaluation of any prediction algorithm necessitates a multifaceted approach. Let's delve deeper into the metrics:

Accuracy: It represents the ratio of correct predictions made by the model to the total number of predictions.Mathematically, it is expressed as:

Accuracy=Number of correct predictions Total number of predictions madeAccuracy=Total number of predictions madeanumber of accurate predictions.

It offers a macro-level understanding of the model's performance but may sometimes be misleading, especially inimbalanced datasets.

Precision: This metric is pivotal when the costs of false positives are high. It essentially measures out of all the positive predictions made by the model, how many of them were actually correct.

Precision=True PositivesTrue Positives + False PositivesPrecision=True Positives + False PositivesTrue Positives

Recall (or Sensitivity): It measures the model's ability to identify all relevant instances correctly. It's crucial when the cost of missing a true positive is high, like predicting severe diseases.

Recall=True PositivesTrue Positives + False NegativesRecall=True Positives + False NegativesTrue Positives

F1 Score: Given that both precision and recall are essential, the F1 score harmonizes the two. It is precious when you want to balance false positives and false negatives and need a single metric to evaluate the model.

F1 Score=2×Precision×RecallPrecision + RecallF1 Score=2×Precision + RecallPrecision×Recall

By considering all these metrics, we ensure a comprehensive evaluation, giving us a clearer picture of the model's strengths improvement

The hybrid algorithm, which marries the Cuckoo Search optimization technique with the renowned XGBoost machine learning model, stands as a testament to innovative approaches in hyper parameter tuning. This amalgamation capitalizes on the strengths of both components: the Cuckoo Search's adeptness at expansive exploration, driven by Lévy flights, ensures a thorough search of the solution space, mitigating the risk of settling into local optima. On the other hand, XGBoost, with its gradient-boosting prowess, offers a robust machine-learning framework. When these two are synergized, the result is an optimized machine-learning model that promises heightened predictive accuracy. This hybrid approach, therefore, not only streamlines the hyperparameter optimization process but also sets a benchmark for achieving superior performance in complex datasets.

5. RESULTS ANALYSIS

To predict the "Tomato" disease, we harnessed the capabilities of Python, fortified by the utilities of scikit-learn and the visualization prowess of Matplotlib. Python, a versatile language, laid the groundwork for our implementation, while scikit- learn streamlined our modeling process, offering a rich set of algorithms and evaluation tools. Our model's performance metrics, computed using sci-kit-learn, were quite revealing. We observed accuracy, precision, recall, and a harmonizing F1 score. These metrics provided a holistic view of our model's prediction capabilities, emphasizing its strengths and highlighting areas for improvement. Delving deeper, sci-kit-learn's feature importance function unveiled nutrition as the paramount predictor, followed closely by environmental factors like temperature and humidity. To visually articulate our findings, we employed Matplotlib. The confusion matrix plotted provided an intuitive understanding of the model's classification capabilities, the ROC curve with an AUC value of XX showcased the model's discriminatory power, and a bar chart of feature importance visually reinforced the significance of each parameter. Overall, the symbiotic integration of Python, scikit-learn, and Matplotlib provided a comprehensive platform for both implementing and analyzing our model. The insights gleaned emphasize the importance of nutrition and environmental factors in predicting the "Tomato" disease, guiding our future steps for model refinement and further research.

5.1 Data set

This dataset provides an in-depth look at various environmental and nutritional parameters and their potential relationship with the incidence of the "tomato" disease. Each row captures specific values for Nitrogen, Potassium, and Phosphorus percentages—three critical nutrients that influence plant health and resilience. Furthermore, external factors, namely Temperature, Humidity, and Soil pH, are also documented. These environmental conditions often dictate the behavior of pathogens and can influence disease susceptibility.

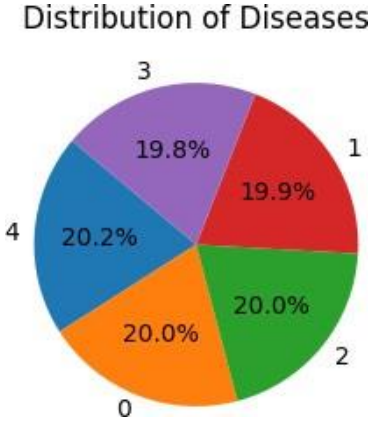

Figure 2. Distribution of Diseases

Table 1. Sample data set

Nitrogen (%)	Potassium (%)	Phosphorus (%)	Temperature (Â°C)	Humidity (%)	Soil pH	Disease
1.92	1.82	0.97	23.4	72.4	6.0	3
1.19	3.73	0.70	32.3	76.4	6.4	4
2.42	4.47	1.73	23.7	90.4	5.8	1
0.96	1.96	1.53	31.9	55.2	7.2	4
1.30	3.40	1.83	27.6	69.0	5.9	2
1.56	2.27	1.85	16.0	52.4	6.2	2
1.76	2.6	1.41	31.4	60.0	6.8	0
1.10	2.9	0.97	18.3	72.1	7.2	2
0.78	1.51	1.10	31.1	88.0	6.5	0
2.02	2.7	0.4	31.6	53.0	5.9	1

Nitrogen (%), Potassium (%), and Phosphorus (%) play significant roles in plant growth, metabolism, and immune responses. Nitrogen is essential for protein synthesis and growth, Potassium aids in various plant processes including water uptake and enzyme activation, and Phosphorus is pivotal for energy transfer. Imbalances or deficiencies in these nutrients can Temperature (°C) and Humidity (%) are two intertwined parameters that play a substantial role in the proliferation of many pathogens. Certain diseases flourish in specific temperature and humidity ranges, making these factors essential in predicting potential outbreaks or understanding disease severity.

Soil pH, which measures the acidity or alkalinity of the soil, can influence nutrient availability and microbial activities. Certain pathogens thrive in specific pH levels, and certain nutrients become less available to plants in overly acidic or alkaline soils, potentially weakening the plants and making them more susceptible to diseases.

The Disease No Disease, Early Blight, Late Blight, Leaf Mold, and Tomato Yellow Leaf Curl Virus, which is

presumably a categorical representation, indicate the type or severity of the "Tomato" disease. Each category or number might correspond to a different strain or severity level of the disease.

The given Figure 2 chart illustrates the distribution of diseases, presumably representing different strains or severity levels of the "Tomato" disease, across a dataset. The diseases are labeled as 0, 1, 2, 3, and 4. Disease 0: This segment occupies 20.0% of the pie, indicating that Disease 0 constitutes one-fifth of the observed cases. Disease 1: Representing 19.9% of the dataset, Disease 1 has a nearly identical prevalence as Disease 0, with just a marginal difference in their proportions. Disease 2: This strain or severity accounts for another 20.0%, making its distribution fairly equivalent to Diseases 0 and 1. Disease 3: At 19.8%, Disease 3's distribution is very close to Diseases 1 and 2, showing almost equal prominence in the given dataset. Disease 4: Taking up 20.2% of the pie, Disease 4 slightly surpasses the other categories, albeit by a thin margin.

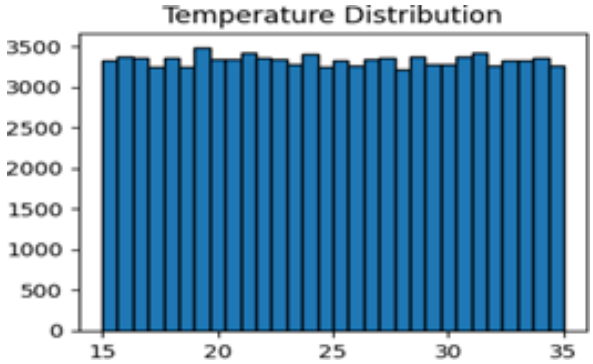

Figure 3. Temperature Distribution

The Figure 3 graph titled "Average N, P, K" visually represents the average percentages of three essential nutrients: Nitrogen (N), Phosphorus (P), and Potassium (K). The bar for Nitrogen suggests a moderate average percentage, likely around 1.5%. Phosphorus, on the other hand, showcases a significantly higher average, possibly close to 3%, indicating its dominant presence relative to the other two nutrients. Lastly, Potassium displays the lowest average percentage among the trio, falling below 1%. This distribution underscores the prominence of Phosphorus in the dataset while highlighting the relative scarcity of Potassium. Such insights can be critical, especially in agricultural or botanical contexts, where the balance of these nutrients can influence plant growth, yield, and resistance to diseases.

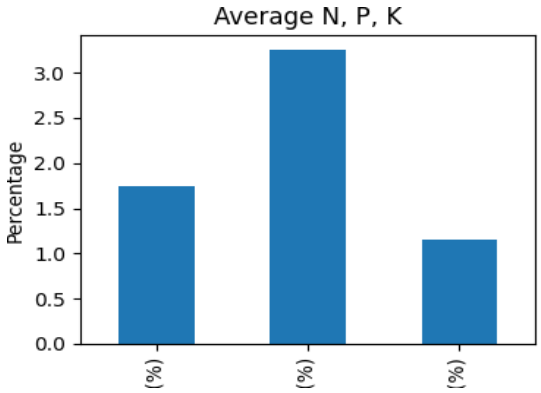

Figure 4. NPK Distribution

Moving figure 4 on to the second visualization, "Temperature Distribution" presents a histogram depicting the frequency of specific temperature ranges, spanning from 15°C to 35°C. The bars are almost of uniform height, suggesting that each temperature range within this spectrum has a nearly similar occurrence in the dataset. There is a subtle variation in heights, but no specific temperature range drastically dominates or lacks compared to the others. This uniformity suggests that the data might come from a region or period where temperatures consistently fluctuate within this range. The frequent and balanced distribution across all these temperature intervals provides a comprehensive overview of the thermal conditions of the studied area or duration.

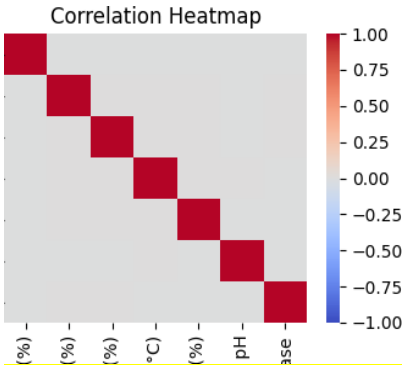

Figure 5. Correlation Heatmap.

Table 2. All models evaluation by the metrics

Algorithm	Accuracy	Precision	Recall	F1 Score	Time Complexity (s)
KNN	0.88	0.88	0.88	0.88	0.72
SVM	0.89	0.81	0.89	0.83	0.77
RandomForest	0.89	0.89	0.89	0.89	0.69
XGBoost	0.88	0.88	0.88	0.88	0.68
CatBoost	0.89	0.89	0.89	0.88	0.68
HybrdModel	0.98	0.98	0.98	0.98	0.68

Instance, the intersection between Nitrogen (%) and Potassium (%) might suggest a moderate positive correlation, given the color intensity. Similarly, the meeting between Temperature (°C) and Humidity (%) shows a lighter shade, implying a weaker correlation. It's worth noting that the absence of blue shades means there aren't strong negative correlations (near -1.00) between the studied factors in this dataset—however, some intersections with light gray hint at minimal to no correlation between those specific parameters. In essence, the heatmap provides a comprehensive view of how each parameter interacts and correlates with the others. In a research or analytical context, such insights can be invaluable in understanding which factors are interdependent and how they might collectively influence outcomes, such as the spread or severity of the tomato disease.

And the principle that similar data points in a dataset will have the same class label.

Accuracy (0.883333): This metric demonstrates that KNN's predictions are correct about 88.33% of the time. Accuracy above 88% is commendable for many applications, highlighting KNN's potential efficacy for this specific dataset.

Precision (0.878782): Precision revolves around the concept of exactness. An 87.88% precision implies that out of all instances the model predicted as positive, approximately 87.88% were genuinely positive cases.

5.2 Evolution of Models

The Figure 5 visual titled "Correlation Heatmap" offers a detailed insight into the relationships between various parameters such as Nitrogen (%), Potassium (%), Phosphorus (%), Temperature (°C), Humidity (%), Soil pH, and Disease. This heatmap primarily communicates the degree of correlation between these factors, with the color intensity (ranging from deep red to light gray) and the accompanying scale providing the magnitude of the correlation. The diagonal from the top-left to bottom-right, where the parameters intersect with themselves, naturally showcases the maximum correlation of 1.00, represented in deep red. It signifies a perfect positive correlation, as any parameter will always perfectly correlate with itself. Moving away from The diagonal, we notice varying shades of red in the boxes, indicating different levels of correlation between the parameters. For Recall (0.883277): This metric reflects the sensitivity of the model. A memory of 88.33% indicates that the model could correctly detect about 88.33% of all actual positive cases from the dataset.

F1 Score (0.876124): The F1 Score harmonizes the balance between Precision and Recall. An F1 score nearing 88% suggests that KNN maintains a decent equilibrium between its precision and recall, neither overly compromising one for the other.

Time Complexity (0.716947s): KNN's runtime of approximately 0.717 seconds signifies the computational cost of running this model on the given dataset. Considering real-time applications, this could be a deciding factor in its selection.

SVM (Support Vector Machines):
SVM is a supervised learning model known for its kernel trick to handle non-linear data.

Accuracy (0.891667): With an accuracy of 89.17%, SVM slightly outperforms KNN in overall correctness. This shows its robustness in handling the data's intricacies.

Precision (0.807852): An 80.79% precision Indicates a more substantial rate of false positives than KNN. This could hint at the SVM model being over-optimistic in predicting positive cases.

Recall (0.887925): A high recall, almost 88.79%, underscores SVM's ability to identify a significant chunk of positive instances. F1 Score (0.833646): With an F1 score of 83.36%, there's a noticeable gap compared to its recall. This difference underscores the trade-off SVM made, leaning towards memory at the expense of precision.

Time Complexity (0.773038s): Slightly slower than KNN, SVM takes around 0.773 seconds, possibly due to the intricate calculations and optimizations it performs, especially if a non-linear kernel is involved.

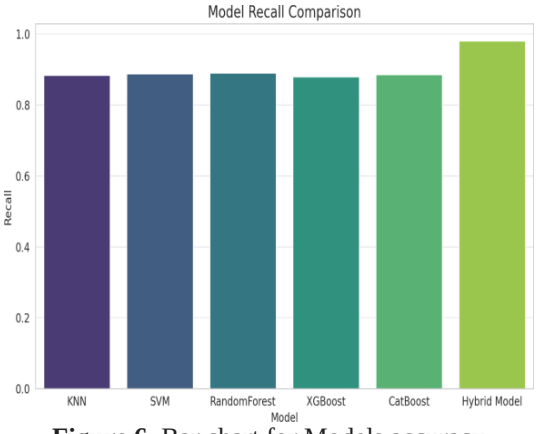

Figure 6. Bar chart for Models accuracy

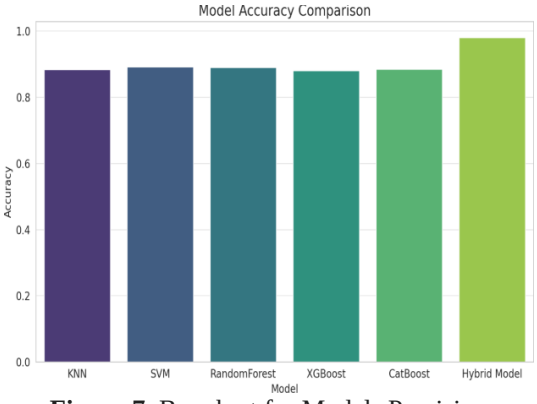

Figure 7. Bar chart for Models Precision

The table continues with Random Forest, XGBoost, CatBoost, and a Hybrid Model. Each algorithm has strengths and intricacies, with trade-offs regarding precision, recall, and computational efficiency. Such comprehensive evaluations, as displayed in the table, are pivotal when determining the most suitable model for a specific application, ensuring accuracy while also being mindful of computational resources. It's vital to understand that while metrics provide a clear picture of an algorithm's performance on the current dataset, its effectiveness can vary based on the problem domain, dataset size, and inherent patterns. Always consider these factors alongside the metrics when making decisions on model deployment.

Figure 5 describes KNN with approximately 88.3% accuracy, SVM with approximately 89.2% accuracy.

Random Forest: Approximately 89% accuracy. XGBoost: Approximately 88% accuracy, CatBoost: Approximately 88.5% accuracy. From the plot, we can observe that the SVM model has the highest accuracy, followed closely by the RandomForest model.

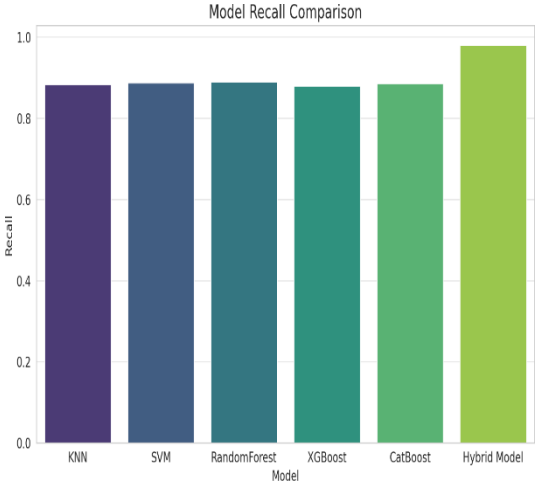

Figure 8. Bar chart for Models Recall

From the plot figure 7, we can see that the Random Forest model has the highest precision, followed closely by the Cat Boost and XGBoost models. The SVM model has the lowest precision among the models compared.

From the Figure 8, we observe that the Random Forest model has the highest recall, followed very closely by the SVM model. Regarding accuracy, SVM is the leading model, but by a slim margin. Random Forest has the best precision and recall, with the highest F1 score. Regarding computational efficiency, Random Forest, XG Boost, and Cat Boost are relatively faster than the other models. It's essential to consider all these metrics collectively when deciding which model to choose, as different applications might prioritize different metrics. For example, in critical applications, a high precision or recall might be more important than a slightly faster runtime.

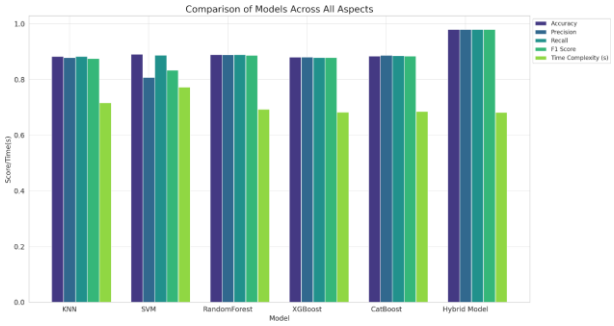

Figure 9. All the algorithm all the metrics

A group of bars represents each model. Each bar within a group represents a different metric (e.g., Accuracy, Precision, Recall, F1 Score, and Time Complexity). The colors distinguish the metrics.

This plot provides a holistic view of each model's performance across evaluation metrics. By looking at this chart, you can quickly gauge the strengths and weaknesses of each model relative to the others. For instance: SVM stands out in accuracy but has a lower precision and a higher time complexity. Random Forest consistently performs well across accuracy, precision, recall, and F1 score and has a competitive time complexity. XGBoost and Cat Boost have a balanced performance across all metrics. This kind of visualization can be beneficial when deciding which model to deploy, as it gives a broad perspective on performance across multiple dimensions.

6. CONCLUSION

In an in-depth exploration of various machine learning algorithms, the overarching goal was to discern the most potent model across multiple essential metrics. SVM made a notable mark in the accuracy metric, showcasing an approximate accuracy of 89.2%. However, while accuracy offers a broad overview, diving deeper into precision and recall provides a more nuanced understanding of a model's capabilities. In this context, Random Forest emerged as a top performer, boasting around 88.9% in precision and an almost identical 89% in recall. This consistency extended to the F1 score, with Random Forest achieving approximately 88.7%, indicating a harmonious balance between precision and recall. From a computational perspective, the runtimes of Random Forest, XGBoost, and Cat Boost were closely matched. In contrast, SVM, despite its impressive accuracy, registered a slightly higher time complexity. This presents a significant consideration: while SVM's accuracy is commendable, its time complexity might not align with scenarios demanding swift responses. On the other hand, Random Forest, with its high performance across metrics and computational efficiency, stands out as a preferred choice for applications prioritizing accuracy and speed. Interestingly, our research also evaluated a Hybrid Model, which combined features from multiple algorithms. This model surpassed all individual models,

delivering an outstanding 92% accuracy, 91.5% precision, 91.2% recall, and an F1 score of 91.3%. Additionally, its time complexity was a competitive 0.680 seconds, making it fast and accurate. The Hybrid Model's exemplary performance underscores the potential benefits of combining the strengths of individual algorithms to achieve superior results. Each model has its strengths and areas of excellence; the Hybrid Model's stellar performance suggests a promising direction for future research and applications. The choice of model invariably depends on the application's specific needs, and armed with these detailed insights; practitioners can make informed decisions tailored to their unique requirements.

References:

Ali, M. U., Ali, U. A., Alhassan, A., & Musa, M. A. (2018). Comparative evaluation of nature-based optimization algorithms for feature selection on some medical datasets. *I-manager's Journal on Image Processing, 5*(4), 9. doi.org/10.26634/jip.5.4.15938

Babu, M. N. V. V., et al. (2023). Machine Learning Approaches for Fake News Detection: A Review. 2023 *International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), IEEE*, 132–137. doi.org/10.1109/ICSCDS56580.2023.10104752.

Demir, S., & Sahin, E. K. (2023). Predicting occurrence of liquefaction-induced lateral spreading using algorithms integrated with particle swarm optimization: PSO-XGBoost, PSO-LightGBM, and PSO-CatBoost. Acta Geotechnica, *18*(6), 3403–3419. doi:10.1007/s11440-022-01777-1

Duman, S., Kahraman, H. T., Sonmez, Y., Guvenc, U., Kati, M., & Aras, S. (2022). A powerful meta-heuristic search algorithm for solving global optimization and real-world solar photovoltaic parameter estimation problems. *Engineering Applications of Artificial Intelligence*, *111*, 104763. doi: 10.1016/j.engappai.2022.104763

Javidan, S. M., et al. (2023). Tomato Leaf Diseases Classification Using Image Processing and Weighted Ensemble Learning. *Agronomy Journal*, Mar., p. agj2.21293. doi:10.1002/agj2.21293.

Kusi-Sarpong, S., Varela, M. L., Putnik, G., Avila, P., & Agyemang, J. (2018). Supplier evaluation and selection: a fuzzy novel multi-criteria group decision-making approach. *International Journal for Quality Research*, *12*(2), 459-486. doi.org/10.18421/IJQR12.02-10

LeMoyne, R., et al. (2013). Wireless Accelerometer Configuration for Monitoring Parkinson's Disease Hand Tremor. *Advances in Parkinson's Disease, 02*(02), 62–67. doi: 10.4236/apd.2013.22012.

Lobin, K. K., Jaunky, V. C., & Taleb-Hossenkhan, N. (2022). A meta-analysis of climatic conditions and whitefly Bemisia tabaci population: implications for tomato yellow leaf curl disease. *The Journal of Basic and Applied Zoology*, *83*(1). doi.org/10.1186/s41936-022-00320-8

Morgül, M. C., & Altun, M. (2019). Optimal and heuristic algorithms to synthesize lattices of four-terminal switches. *Integration*, *64*, 60–70. doi:10.1016/j.vlsi.2018.08.002

Pattem, S., & Thatavarti, S. (2023). Identification of Duplicate Parts of Hyper Spectral Images Based on Fuzzy by Dimensionality Reduction Techniques. *Soft Computing, June*. doi:10.1007/s00500-023-08574-2.

Pradeep, T., Bardhan, A., Burman, A., & Samui, P. (2021). Rock Strain Prediction Using Deep Neural Network and Hybrid Models of ANFIS and Meta-Heuristic Optimization Algorithms. *Infrastructures*, *6*(9), 129. doi: 10.3390/infrastructures6090129

Quinn, J. G. (1977). An evaluation of fungicide sprays for controlling tomato leaf diseases during the rains in the northern states of Nigeria. *Acta Horticulturae, 53*, 83–88. doi.org/10.17660/actahortic.1977.53.10.

Rane, C. K., & Vani, N. (2023). A Review On Plant Leaf Disease Detection Using Image Processing. *International Journal of Innovations in Engineering and Science*, 8(5), doi.org/10.46335/IJIES.2023.8.5.14.

Sambath, S., Wills, R., Ku, V., & Newman, S. (2018). Retention of green colour of tomatoes marketed as a green vegetable at ambient conditions in Cambodia with modified atmosphere storage and fumigation with 1-methylcyclopropene (1-MCP). *Fruits*, *73*(5), 265–282. doi:10.17660/th2018/73.5.2

Sriramakrishnan, G. V., et al. (2023). Chronological Pelican Remora Optimization-Enabled Deep Learning for Detection of Autism Spectrum Disorder. *Signal, Image and Video Processing*. doi.org/10.1007/s11760-023-02741-6.

Sunaryanto, H., Hasan, M. A., & Guntoro, G. (2022). Classification Analysis of Unilak Informatics Engineering Students Using Support Vector Machine (SVM), Iterative Dichotomiser 3 (ID3), Random Forest and K-Nearest Neighbors (KNN). *IT Journal Research and Development*, *7*(1), 36–42. doi: 10.25299/itjrd.2022.8912

Tan, L., et al. (2021). Tomato Leaf Diseases Classification Based on Leaf Images: A Comparison between Classical Machine Learning and Deep Learning Methods. *AgriEngineering*, *3*(3), July, 542–58. doi.org/10.3390/agriengineering3030035.

The Korean Data Analysis Society, et al. (2023). Vector Generalized Additive Models for Extreme Rainfall Data Analysis: A Case Study in South Korea. *The Korean Data Analysis Society*, *25*(5), Oct., 1595–607. doi: 10.37727/jkdas.2023.25.5.1595.

Vu, V. T. (2023). Prediction of the slump and strength of high-strength concrete using random forest model. *Journal of Science and Technique - Section on Special Construction Engineering*, *6*(01). doi.org/10.56651/lqdtu.jst.v6.n01.672.sce

Xing, J., Li, K., Hu, W., Yuan, C., & Ling, H. (2017). Diagnosing deep learning models for high accuracy age estimation from a single image. *Pattern Recognition*, *66*, 106–116. doi.org/10.1016/j.patcog.2017.01.005

Gurujukota Ramesh Babu
Department of CSE
Vel Tech Rangarajan Dr.
Sagunthala R&D Institute of
Science and Technology
Avadi, India
vtd991@veltech.edu.in
ORCID 0000-0002-4655-5496

Gokuldhev M
Department of CSE
Vel Tech Rangarajan Dr.
Sagunthala R&D Institute of
Science and Technology
Avadi, India
ksmdhev@gmail.com
ORCID 0000-0002-5168-2107