RESEARCH ARTICLE OPEN ACCESS

Graph-Based Disease Prevention for Sustainable Avocado Farming: A Knowledge Graph Approach

Neha S*, Abseena Habeeb**
*School of Data Analytics, Mahatma Gandhi University, Kottayam, Kerala, India

Abstract:
Avocado is a valuable crop that is prone to a number of diseases brought on by environmental factors and pathogens. Disease management and prevention depend on an understanding of the connections between avocado species, ill-nesses, and their causes. This study models and visualizes these interactions by building a knowledge graph using network analysis techniques. Disease con-nectivity, common causative factors, and vulnerable species are highlighted in the knowledge graph. We locate important disease hubs and patterns of dis-ease spread by using graph visualization techniques, community detection, and centrality measures. The results help with early detection and targeted dis-ease management by providing farmers and agricultural researchers with useful information.

Keywords: Knowledge Graph, Data-Driven Disease Prevention, Plant Disease Surveillance, Smart Farming.

1. Introduction

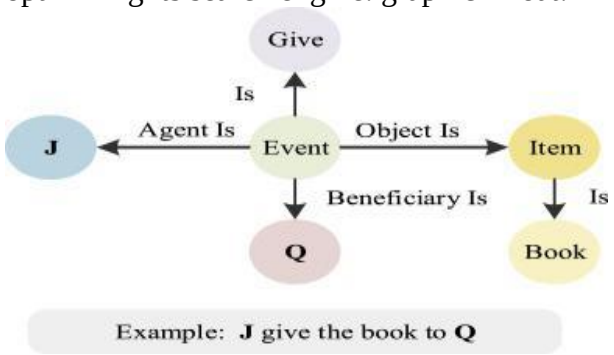
The The avocado(Persea americana Mill.) is an evergreen tree in the laurel family (Lauraceae)[1]. It is native to the Americas and was first domesticated in Mesoamerica more than 5,000 years ago. It was prized for its large and unusually oily fruit. The tree likely originated in the highlands bridging south-central Mexico and Guatemala[2]. Avocado trees have a native growth range from Mexico to Costa Rica[3].

Its fruit, is botanically a large berry containing a single large seed[4].Sequencing of its genome showed that the evolution of avocados was shaped by polyploidy events and that commercial varieties have a hybrid origin[2]. Avocado trees are partly self-pollinating, and are often propagated through grafting to maintain consistent fruit output. Avocados are presently cultivated in the tropical and Mediterranean climates of many countries. As of 2023, Mexico is the world's leading producer of avocados, supplying 29 Plant diseases pose a serious threat to the yield and quality of avocados, making cultivation extremely difficult. The intricate relationships between various species, pathogens, and environmental factors are frequently missed by traditional disease analysis techniques. These relationships can be represented in an organized manner with a knowledge graph, which improves analysis and visualization. In order to analyze avocado diseases, this study builds a knowledge graph with the following goals in mid: • Identify key disease hubs that impact several species; • Identify common causes that lead to multiple diseases. • Find trends in the vulnerability and connectivity of diseases.

1.1 Knowledge Graph

The knowledge graph is a large-scale semantic network, as shown in Fig. 1[5]. It is a graph-based data structure, which consists of nodes and edges. "Entities" are repre-sented by nodes, and each bite is a "relationship" between entities[6]. Things in the real world, such as people, animals, companies, place names, and telephones, can be represented by entities. A relation can express a specific connection between different entities. It

originated from Google's purpose of optimizing its search engine. graphicx float.



2. Related Work

Knowledge graphs (KGs) have emerged as powerful tools for organizing agricultural information. Recent work by Zhao et al. [7] demonstrated the effectiveness of graph-based approaches for modeling crop disease networks, though their framework didn't explicitly represent causal relationships. The Planteome project [8] developed com-prehensive plant ontologies, but focused primarily on taxonomic relationships rather than disease mechanisms.

For visualization techniques, Bastian et al. [9] introduced Gephi, which has become a standard for network analysis, though domain-specific adaptations remain necessary .

Recent advances in KG applications for precision agriculture [11] have shown promising results in disease prediction, but few studies have focused specifically on tree crops like avocado.

Our work builds upon these foundations while addressing three key gaps: (1) explicit causal modeling in agricultural KGs, (2) domain-optimized visualization for tree crop diseases, and (3) integrated analytical metrics for disease impact assessment.

3. Methodology

3.1 Data Collection

The dataset used in this study was collected from field reports, agricultural research papers, and disease databases. It included three main types of information: diseases (Phytophthora root rot, Algal Leaf Spot, and Anthracnose), causes(algal,fungal,mites ,etc) and avocado species (Persea americana, Persea schiedeana, etc.). The dataset was in the form of excel file and was uploaded, and pandas in Python was used to process it. The analysis involved cleaning the dataset to remove any duplicates or inconsistencies in it , ensuring the reliability of the results. Significant associations between the avocado species and their vulnerability to different diseases under various environmental conditions were found through subsequent statistical analyses.

3.2 Data Preprocessing

To guarantee uniformity, consistency, and readability, data preprocessing was done prior to building the knowledge graph[12]. Pandas was initially used to clean and format the dataset, removing any extraneous spaces and irregularities[13]. Underscores were used in place of spaces in the names of the species, disease, and cause in order to avoid processing errors within the graph. Before building the entire knowledge graph, a subset of five records was also first extracted and examined to confirm relationships and guarantee the correctness of entity connections [14].

3.3 Knowledge Graph Construction

In order to efficiently model the relationships between avocado species, diseases, and their causes, the knowledge graph was dynamically created using NetworkX[15], a Python library for network analysis. Three main node categories made up the graph structure: species (blue nodes) represented various avocado varieties, diseases (orange nodes) indicated different plant infections, and causes (green nodes) included envi-ronmental factors and pathogens. Directed edges were used to create relationships; the 'Affectedby' edge connected avocado species to the diseases to which they are vulnerable, and the 'Causedby' edge linked diseases to the pathogens or environmen-tal stressors to which they are related. To guarantee directionality and accurately depict the pathways by which diseases are transmitted, a directed graph (DiGraph) was employed.

3.4 Graph Visualization Techniques

Matplotlib and NetworkX were used to visualise the graph with a number of enhance-

ments to help interpret disease relationships: color-coded nodes (orange for diseases, green for pathogens, and blue for avocado species); custom edge labels explicitly defined relationships ('affectedby', 'causedby'); arrow styling ensured that the direc-tion of disease transmission was clear; and a custom legend was added using Matplotlib to make the graph easier to read and interpret.

3.5 Small Subset vs Full Graph Analysis

A stepwise approach was followed to scale up the analysis, ensuring the structural integrity of the knowledge graph. Initially, a small subset of five records was visualized using a spring layout to validate node connections. Once confirmed, the entire dataset was processed with a shell layout, optimizing for large-scale visualization and efficiency. This method allowed for early error detection and ensured the graph was structurally sound before full-scale processing.

3.6 Graph Analysis and Network Metrics

We employed several network science methods to extract insights from the knowledge graph, following established analytical frameworks in agricultural network analysis . Betweenness centrality [16] identified critical bridge nodes serving as disease trans-mission points, while degree centrality [17] revealed high-risk disease hubs affecting multiple avocado species. The Louvain algorithm [18] detected community structures, clustering biologically related diseases and species. Pathway analysis [19] traced dis-ease transmission routes across avocado cultivars. Together, these metrics provided actionable intelligence about disease susceptibility patterns and potential intervention targets for improved phytosanitary management.

3.7 Graph Export and Future Applications

For reference and documentation, the entire graph was saved as a PNG image.For farmers, policymakers, and agricultural researchers, this structured knowledge graph is an invaluable tool for comprehending disease relationships. In the future, the dataset might be expanded to include more avocado species and environmental factors. It might also be updated dynamically with real-time disease reports to ensure more precise and timely disease monitoring.

3.8 Evaluation and Validation

Plant pathologists were consulted to verify the biological significance of disease rela-tionships, and the results were cross-validated with agricultural disease reports to guarantee the precision and dependability of the graph-based disease model. In order to improve disease prediction and the model's efficacy in tracking and treating avocado diseases, future research will concentrate on incorporating real-time disease data.

4. Results and Discussion

The avocado species and disease knowledge graph was constructed using NetworkX, representing the relationships between avocado species, diseases, and their causes. The graph consists of three nodes they are avocado species (Persea americana, Persea schiedeana,Persea floccosa ,etc), diseases (Algal Leaf Spot, Anthracnose, Cercospora spot,Scab), and causal factors (Algal, mites, Fungal). Visualization was done using spring and shell layouts, ensuring clear representation of disease interactions. Nodes were colored, with species in blue, diseases in orange, and causes in green. The result-ing knowledge graph effectively maps the connections between species and disease , providing a structured understanding of disease transmission.

The analysis of degree centrality revealed that Phytophthora root rot and Anthrac-nose are the most connected diseases, affecting multiple avocado species. This indicates that these diseases are major threats to avocado farming and require early detection and targeted management strategies. Managing high-centrality diseases and species is crucial for preventing large-scale outbreaks.

community detection using the Louvain algorithm grouped diseases into distinct clusters, highlighting common transmission patterns. Fungal diseases such as Anthrac-nose and Cercospora spot formed a single cluster, suggesting shared environmental factors and

potential cross-disease intervention strategies. Pest-related issues, includ-ing Mites and Algal Leaf Spot, were grouped together, reinforcing the need for integrated pest management solutions. The identification of disease clusters indi-cates that controlling one disease in a cluster could reduce the prevalence of multiple infections, improving overall avocado health.

Directed Graph (DiGraph) in NetworkX .It Builds a directed knowledge graph of avocado species, diseases, and causes.Spring Layout used for a subset graph to optimize node spacing.Shell Layout is used for full graph visualization to structure large-scale data.Color Legends improves readability by categorizing species, diseases, and causes. Node and Edge Relationships defines affected by and caused by relationships between species, diseases, and causes.

By analyzing the knowledge graph that are constructed we can identify high risk diseases and their causes.. Diseases with high degree centrality, such as Phytophthora root rot (centrality score: 0.42) and Anthracnose (0.38), are likely to remain persistent threats requiring preventive measures. If an avocado species shares multiple connections with known affected species, it is likely at risk of future infection. For example, if Fusarium wilt is strongly linked to three different avocado species in the graph, it suggests a high probability of spreading to related varieties. Additionally, diseases that share common causal factors such as Botryosphaeria dieback and Cercospora spot, both linked to fungal pathogens indicate that controlling one pathogen could prevent multiple diseases. Community detection algorithms reveal clusters of related diseases. If a newly identified disease appears within an existing disease cluster, it is likely to share transmission methods and treatment responses with other diseases in that group. For example, the graph may show that Avocado sunblotch clusters with Scab and Black spot, indicating possible similarities in disease progression. The knowledge graph provides a structured visualization of disease interactions, reveal-ing critical insights for avocado disease management. The ability to identify major disease hubs and shared causes enables targeted interventions. These findings can aid researchers, agronomists, and farmers in: • Developing early warning systems for high-risk diseases. • Prioritizing disease-resistant avocado species. • Implementing preventive measures for frequently occurring pathogens. One limitation of this study is the reliance on existing disease databases, which may not account for emerging dis-eases. Future work could integrate real-time field data, genetic factors, and climatic conditions to enhance predictive capabilities.

4.1 Avocado Images

4.2 Diseases

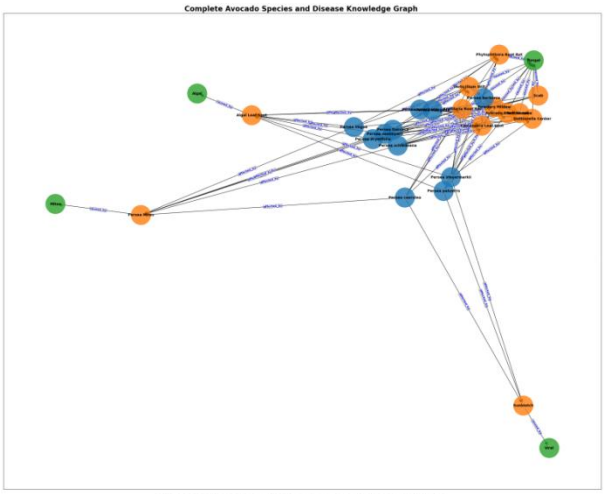
Algal leaf spot

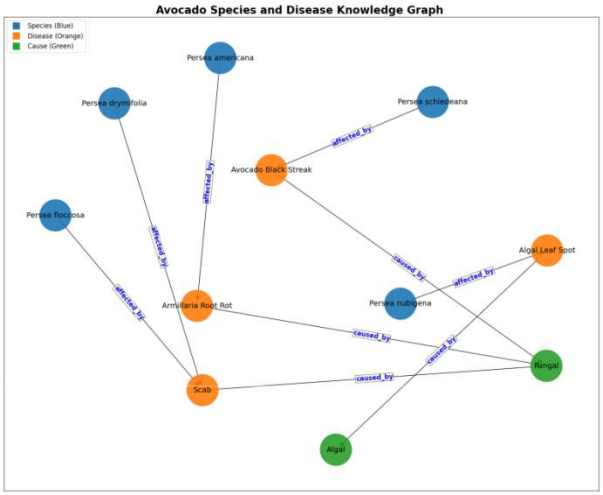
Scab

Anthracnose

Persea Mites

4.3 Knowledge Graphs





5. Conclusion

Relationships between species, pathogens, and environmental factors are the main focus of this study, which presents a graph-based knowledge framework for comprehending disease interactions in avocado farming [11]. By identifying high-risk avocado species and important disease hubs through network analysis, we were able to uncover common causative agents and patterns of disease transmission [18]. In order to help with targeted disease prevention, the most influential diseases were identified through the analysis of centrality measures . According to the findings, specific pathogens and environmental factors play a crucial role in the spread of disease, offering farmers and agricultural researchers important new information . This strategy uses graph-based disease modelling to help with early detection, risk assessment, and well-informed decision-

making, which promotes more sustainable avocado farming [20]. This study could be expanded in the future by integrating machine learning models for automated risk prediction, real-time disease monitoring, and climate-based interventions [21].

References

[1]. Galindo-Tovar, M. E., Ogata-Aguilar, N., Arzate-Fern´andez, A. M. (2008). Some aspects of avocado (Persea americana Mill.) diversity and domestication in Mesoamerica. Genetic Resources and Crop Evolution, 55, 441-450.

[2]. Rend´on-Anaya, M., Ibarra-Laclette, E., M´endez-Bravo, A., Lan, T., Zheng, C., Carretero-Paulet, L., ... Herrera-Estrella, L. (2019). The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. Proceedings of the National Academy of Sciences, 116(34), 17081-17089.

[3]. Silva, S. S. D., Sim˜ao-Bianchini, R., Sim˜oes, A. R. G., Costea, M. (2021). Disentangling parasitic vines in the tropics: taxonomic notes for an accurate identification of Cuscuta (Convolvulaceae) and Cassytha (Lauraceae). Rodrigu´esia, 72, e01062020.

[4]. Yahia, E. M., Woolf, A. B. (2011). Avocado (Persea americana Mill.). In Postharvest biology and technology of tropical and subtropical fruits (pp. 125-186e). Woodhead Publishing.

[5]. Zhu, D., Xie, L., Chen, B., Tan, J., Deng, R., Zheng, Y., ... Ip, A. W. (2023). Knowledge graph and deep learning based pest detection and identification system for fruit quality. Internet of Things, 21, 100649.

[6]. Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E. (2015). A review of relational machine learning for knowledge graphs. Proceedings of the IEEE, 104(1), 11-33.

[7]. Chen, G., Auerswald, K. (2018). A user-friendly calculator for determining oxygen isotope composition of body water in cows based on the MK model and examples of its applications. Computers and Electronics in Agriculture, 154, 248-255.

[8]. Cooper, L., Meier, A., Laporte, M. A., Elser, J. L., Mungall, C., Sinn, B. T., Jaiswal, P. (2018). The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic acids research, 46(D1), D1168-D1180.

[9]. Bastian, M., Heymann, S., Jacomy, M. (2009, March). Gephi: an open source software for exploring and manipulating networks. In Proceedings of the international AAAI conference on web and social media (Vol. 3, No. 1, pp. 361-362).

[10]. Munz, J., Gindele, N., Doluschitz, R. (2020). Exploring the characteristics and utilisation of Farm Management Information Systems (FMIS) in Germany. Computers and Electronics in Agriculture, 170, 105246.

[11]. Kamilaris, A., Prenafeta-Bold´u, F. X. (2018). Deep learning in agriculture: A survey. Computers and electronics in agriculture, 147, 70-90.

[12]. Garc´ıa, S., et al. (2016). ”Data preprocessing in data mining.” Springer International Publishing.

[13]. McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. Python for high performance and scientific computing, 14(9), 1-9.

[14]. Paulheim, H. (2016). Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic web, 8(3), 489-508.

[15]. Hagberg, A., Swart, P. J., Schult, D. A. (2008). Exploring network structure, dynamics, and function using NetworkX (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Laboratory (LANL), Los Alamos, NM (United States).

[16]. Brandes, U. (2008). On variants of shortest-path betweenness centrality and their generic computation. Social networks, 30(2), 136-145.
[17]. Freeman, L. C. (2002). Centrality in social networks: Conceptual clarification. Social network: critical concepts in sociology. Londres: Routledge, 1(3), 238-263.
[18]. Blondel, V. D., Guillaume, J. L., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), P10008.
[19]. Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., ... Bagos, P. G. (2011). Using graph theory to analyze biological networks. BioData mining, 4, 1-27.
[20]. Nelson, R. (2020). International plant pathology: past and future contributions to global food security. Phytopathology, 110(2), 245-253.
[21]. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674.
[22]. Pradeep Jha, Deepak Dembla, Widhi Dubey, "Implementation of Machine Learning Classification Algorithm Based on Ensemble Learning for Detection of Vegetable Crops Disease", International Journal of Advanced Computer Science & Applications, Vol. 15, Issue. 1, 2024.
[23]. Pradeep Jha, Deepak Dembla, Widhi Dubey, "Implementation of Transfer Learning Based Ensemble Model using Image Processing for Detection of Potato and Bell Pepper Leaf Diseases", International Journal of Intelligent Systems and Applications in Engineering, Vol. 12, pp. 69-80, 2024.
[24]. Jha, P., Dembla, D. & Dubey, W. Deep learning models for enhancing potato leaf disease prediction: Implementation of transfer learning based stacking ensemble model. Multimed Tools Appl 83, 37839–37858 (2024).
[25]. P. Jha, D. Dembla and W. Dubey, "Comparative Analysis of Crop Diseases Detection Using Machine Learning Algorithm," 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), pp. 569-574, 2023.
[26]. Jha, P., Dembla, D., Dubey, W., "Crop Disease Detection and Classification Using Deep Learning-Based Classifier Algorithm", Emerging Trends in Expert Applications and Security. ICETEAS 2023. Lecture Notes in Networks and Systems, vol 682. 2023.