A Comparative Analysis of Decision Tree and XGboost Algorithms for Predicting Type 2 Diabetes Among Patients

Velasco, Roz Guiller Andre

*Bachelor of Science in Computer Science, National University Lipa City Batangas, Philippines*

[velascorf@students.nu-lipa.edu.ph](mailto:velascorf@students.nu-lipa.edu.ph)

**Abstract- Deep learning, particularly Convolutional Neural Networks (CNNs), has demonstrated remarkable success in image classification tasks across multiple domains, including agriculture. This study explores the effectiveness of three CNN architectures—DenseNet121, EfficientNetV2S, and MobileNetV2—for binary classification of tomato fruit images into *tomato* and *non-tomato* categories. Utilizing a dataset of 2,000 labeled images, split into 70% training, 25% validation, and 5% testing sets, each model was evaluated in both Non-Fine-Tuned and Fine-Tuned configurations.**

**Results showed that all three models achieved high accuracy, with MobileNetV2 reaching 99% test accuracy post-fine-tuning, DenseNet121 achieving 100%, and EfficientNetV2S maintaining perfect scores across all metrics. Despite EfficientNetV2S being the most accurate, its computational complexity may not be ideal for small-scale applications. In contrast, MobileNetV2 proved to be the most efficient, balancing accuracy and performance, making it highly suitable for deployment in resource-constrained agricultural settings.**

**The findings suggest that for small classification tasks with limited data, lightweight models like MobileNetV2 offer an optimal solution. Future work may include expanding the dataset, testing real-time performance on edge devices, and extending the classification to multiple tomato ripeness stages or defect types to enhance applicability in real-world scenarios.**

**Keywords- Deep Learning, DenseNet121, EfficientNetV2S, MobileNetV2, Tomato Fruit**

1. INTRODUCTION
2. *Supervised Learning in Studies*

Supervised learning is an essential part of machine learning today, as it allows models to discover relationships and patterns in labeled data sets. By designating known output categories or values to input features it allows the algorithms to make accurate predictions. Furthermore, a type of supervised learning algorithm called a decision tree classifier maps patterns to predefined categories by

estimating class labels (Y) from input data (X) (Safavian & Landgrebe, 1991). The skill to categorize data into predefined classes using labeled datasets is useful for classification, which enables them to be used for a variety of classification challenges, particularly in healthcare.

1. *Decision Trees in Healthcare*

Decision Tree algorithms have become valuable tools in healthcare due to their clarity and ability to show decision-making processes in a straightforward, organized way. In classifying data by following a series of rules based on feature values, like patient details. The transparency of decision trees allows them to be useful in decisions in a clinical setting where explainability is essential. In a study of predicting the risk of mortality of critically ill adult COVID-19 patients found that decision trees were effective in prediction by identifying critical predictors like age, comorbidities, and laboratory results. (Alzahrani et al., 2022). Because of the clarity and adaptability of decision tree models, they are an important tool for medical diagnosis, prognosis, and treatment planning.

1. *Diabetes Prediction Importance*

Diabetes mellitus is a chronic and globally prevalent disease and presents major health hazards, such as kidney failure, neuropathy, and cardiovascular diseases. The chance of avoiding the associated comorbidities will rise if it is identified early. Decision tree classifiers can differentiate between diabetic and non-diabetic cases, which would help avoid the comorbidities of diabetes. The main factors that influence the condition are also provided by these models (Shin et al., 2022). A predictive model can detect a disease like diabetes, which helps healthcare professionals in giving preventive or modified treatment plans to their patients.

1. *Objective of the Study*

This research compares Decision Tree and XGBoost, particularly their performance in predicting Type 2 diabetes using an individual’s clinical data. By evaluating the model’s accuracy, precision, recall, and F1-score. The study aims to determine which algorithm provides more reliable and understandable results for early diabetes prediction, consequently showing what is best suited for practical healthcare use and informed medical decision-making.

1. LITERATURE REVIEW

Machine learning has greatly affected healthcare analytics, specifically in the prediction and diagnosis of chronic diseases namely Type 2 Diabetes Mellitus. Because of its interpretability and ease of use in conveying classification logic, the Decision Tree algorithm has gained widespread adoption among other algorithms. For example, Sadiq et al. (2024) used a Decision Tree model to predict diabetes in a medical dataset and found that age, BMI, and glucose were important predictors. They also achieved a high classification accuracy. This study showed that decision trees are a dependable tool for medical decision assistance since they can handle both numerical and categorical data efficiently.

Building upon these developments, in healthcare prediction tasks, one of the most powerful and emerging ensemble learning algorithms is Extreme Gradient Boosting (XGBoost). XGBoost sequentially combines multiple weak learners rather than the traditional Decision Trees, which results in minimized classification errors. In a recent study, Maulana et al. (2023) utilized and fine-tuned XGBoost for diabetes prediction using the Pima Indian Diabetes Dataset. They compared logistic regression and random forest models to the fine-tuned XGBoost, and it got the highest in accuracy, precision, sensitivity, and F1-score. The study’s results also showed that handling imbalanced datasets and complex feature interactions, especially in the medical field, is one of the strengths of XGBoost.

In addition, other researchers have compared Decision Tree and XGBoost to assess their strengths and limitations in clinical settings. Such as, Talebi Moghaddam et al. (2024) to predict diabetes risk among adults, who compared multiple machine learning algorithms, including Decision Tree, Random Forest, and XGBoost. Their findings revealed that the Decision Tree had an advantage for its interpretability and simplicity in medical applications, while XGBoost, in terms of precision and recall, outperformed the other models.

Furthermore, to enhance transparency in clinical decision-making, explainable artificial intelligence (XAI) frameworks have been increasingly integrated with tree-based models. Recent studies, mainly from 2022 to 2025, stressed the importance of the analysis of feature importance in identifying dominant factors such as glucose levels, insulin, BMI, and age in influencing the prediction outcomes of diabetes. This approach validates that the predictions of models remain explainable and current medical knowledge.

Lastly, the continuous comparison of interpretable models like Decision Trees with high-performance ensemble algorithms such as XGBoost provides valuable insight into achieving both accuracy and explainability in healthcare AI systems. These findings support the development of predictive frameworks that not only perform well statistically but also maintain clinical reliability and trust among medical practitioners.

1. METHODOLOGY

This section describes the procedures employed in data collection, as well as the statistical tools and analytical techniques used to process, interpret, and analyze the gathered data.

1. *Models Used*

**Decision Tree.** A supervised learning algorithm widely used for both classification and regression tasks. To maximize information gain or minimize impurity (e.g., Gini index or entropy), a decision tree functions by recursively partitioning data into subsets based on feature values. This procedure results in a simple and interpretable tree format that the model provides. Because of that, Decision Trees are particularly useful in guiding decisions in finance, education, and healthcare. According to Safavian and Landgrebe (1991), decision trees decompose complex decisions into a collection of simpler ones, offering both accuracy and interpretability in predictive modeling.

**Extreme Gradient Boosting (XGBoost).** It is built upon the gradient boosting framework, which sequentially combines multiple weak learners, which are typically decision trees, resulting in minimized classification errors and enhance predictive performance. Due to the incorporation of XGBoost with regularization techniques, unlike traditional Decision Trees, they can prevent overfitting, making them robust for high-dimensional and noisy data.

In numerous healthcare prediction studies, XGBoost has found great success in disease diagnosis and risk assessment because of its efficiency and higher accuracy. Maulana et al. (2023) said that by optimizing model generalization and computational performance, XGBoost is significantly greater than the conventional decision tree.

1. *Application of Models* 
   1. **Data Collection-** The dataset used in this study was obtained from the Healthcare Diabetes Dataset available on Kaggle (Nandita Pore, 2022). It consists of patient medical attributes such as glucose level, blood pressure, BMI, insulin concentration, and age, alongside the binary target variable Outcome, which indicates whether a patient is diagnosed with Type 2 Diabetes (1) or not (0). The dataset contains 2,769 patient records, each representing unique clinical data points necessary for machine learning-based prediction. The data was divided into 80% for training and 20% for testing to evaluate both model learning and generalization performance effectively.

2. **Data Preprocessing**- Before model training, data preprocessing was performed to handle missing and inconsistent values. Columns such as Glucose, BloodPressure, SkinThickness, Insulin, and BMI contained zero entries, which were replaced with NaN and subsequently imputed using the median strategy via the SimpleImputer from sklearn.

Categorical variables were encoded using one-hot encoding, and all numerical attributes were standardized using the StandardScaler to ensure that the features contributed equally to the models’ performance. Finally, the preprocessed dataset was split into feature variables (X) and the target variable (Y), preparing it for training.

Exploratory data analysis (EDA) was conducted using seaborn and matplotlib to visualize class distribution, feature correlation, and the relationship between key predictors and diabetes outcomes. Figure 1 displays the correlation heatmap used to identify the most influential variables.

3. **Model Training**- Two supervised learning algorithms were implemented—Decision Tree and Extreme Gradient Boosting (XGBoost)—to compare their predictive performance.

1. Decision Tree Classifier

The Decision Tree Classifier was implemented using DecisionTreeClassifier from sklearn.tree. The model was trained on the preprocessed training set (x\_train, y\_train) and tested on (x\_test, y\_test). Performance metrics such as accuracy, precision, recall, and F1-score were computed using sklearn.metrics. The model achieved strong results, and the Confusion Matrix confirmed effective classification of diabetic versus non-diabetic cases. Feature importance analysis identified Glucose, BMI, and Age as the top contributing features, visualized via bar charts.

b. Extreme Gradient Boosting (XGBoost) Classifier

The XGBoost Classifier (XGBClassifier) was trained under similar data conditions to ensure comparability. Unlike the Decision Tree, XGBoost sequentially adds weak learners to minimize classification errors, improving accuracy and robustness. Performance metrics were calculated similarly, revealing superior predictive accuracy and recall compared to the Decision Tree model. Feature importance visualization also showed that Glucose and BMI were the strongest indicators of diabetes risk.

4. **Model Validation and Evaluation**- To ensure reliability and generalization, both models underwent 10-fold Stratified Cross-Validation using StratifiedKFold(n\_splits=10). Mean accuracy and standard deviation were calculated across folds for each model. Additionally, ROC-AUC (Receiver Operating Characteristic – Area Under Curve) scores were computed to measure each model’s discriminative ability. Visualization through ROC Curves illustrated that the XGBoost model achieved a higher AUC compared to the Decision Tree, indicating stronger classification performance. A bar plot comparison summarized the mean accuracy across models, confirming XGBoost’s superior performance in predictive accuracy and stability.

5. **Feature Importance and Model Insights**- Feature importance graphs were generated for both models. For the Decision Tree, the most significant predictors were Glucose, BMI, and Age. For XGBoost, similar features ranked highest, with additional emphasis on Insulin and Blood Pressure. These insights align with established medical findings, reinforcing that glucose levels, body composition, and age are the strongest determinants in diabetes prediction.

Both models demonstrated effective predictive capability for Type 2 Diabetes. The Decision Tree offered interpretability and transparency, making it suitable for clinical applications requiring explainable logic. Meanwhile, XGBoost provided enhanced predictive power and generalization through ensemble boosting. The combination of accuracy, recall, and AUC results confirmed XGBoost as the more robust model for healthcare prediction tasks involving structured tabular data.

1. RESULTS AND DISCUSSION
2. CONCLUSIONS
3. RECOMMENDATIONS

REFERENCES

1. *A Hybrid Model using MobileNetv2 and SVM for Enhanced Classification and Prediction of Tomato Leaf Diseases*. (n.d.). ResearchGate.

https://[www.researchgate.net/publication/374099550\_A\_Hybrid\_Mod](http://www.researchgate.net/publication/374099550_A_Hybrid_Mod) el\_using\_MobileNetv2\_and\_SVM\_for\_Enhanced\_Classification\_and

\_Prediction\_of\_Tomato\_Leaf\_Diseases?utm\_source=chatgpt.com

1. Anam, S., Ayu, D. L. C., Deny, T. a. F., Tarno, H., & Fitriah, Z. (2024). New Tomato Leaf Disease Classification Method Based on DenseNet121 with Bat Algorithm Hyperparameters Optimization.

*EBSCOhost*. https://doi.org/10.15849/IJASCA.240730.03

1. Biswal, A. (2025, April 13). *Top 25 deep learning applications used across industries*. Simplilearn.com. https://[www.simplilearn.com/tutorials/deep-learning-tutorial/deep-](http://www.simplilearn.com/tutorials/deep-learning-tutorial/deep-) learning-applications?utm\_source=chatgpt.com

*levels*. (n.d.). ResearchGate. https://[www.researchgate.net/publication/381290420\_CNN\_Algorith](http://www.researchgate.net/publication/381290420_CNN_Algorith) m\_Approach\_for\_Classification\_of\_Tomato\_Fruit\_Maturity\_Levels? utm\_source=chatgpt.com

1. Flatworld Solutions Editorial Team. (2025, April 9). *7 Transformative applications of convolutional neural networks*. Global BPO, BPA, AI Services Company. https://[www.flatworldsolutions.com/data-](http://www.flatworldsolutions.com/data-) science/articles/7-applications-of-convolutional-neural- networks.php?utm\_source=chatgpt.com
2. *Introduction to DenseNet with TensorFlow*. (2020). https://[www.pluralsight.com/resources/blog/guides/introduction-to-](http://www.pluralsight.com/resources/blog/guides/introduction-to-) densenet-with-tensorflow
3. Niu, L. (2024). Research on the identification of tomato leaf diseases based on multi-scale feature fusion in EfficientNetV2-S. In *Lecture notes on data engineering and communications technologies* (pp. 170–183). https://doi.org/10.1007/978-3-031-71619-5\_15
4. *Papers with Code - EfficientNetV2: Smaller Models and Faster Training*. (2021, April 1). https://paperswithcode.com/paper/efficientnetv2-smaller-models-and- faster
5. Phan, Q., Nguyen, V., Lien, C., Duong, T., Hou, M. T., & Le, N. (2023). Classification of tomato fruit using YoLov5 and convolutional neural network models. *Plants*, *12*(4), 790.

https://doi.org/10.3390/plants12040790

1. *Research on the identification of tomato leaf diseases based on multi- scale feature fusion in EfficientNetV2-S*. (n.d.). ResearchGate. https://[www.researchgate.net/publication/384847998\_Research\_on\_th](http://www.researchgate.net/publication/384847998_Research_on_th) e\_Identification\_of\_Tomato\_Leaf\_Diseases\_Based\_on\_Multi- scale\_Feature\_Fusion\_in\_EfficientNetV2-S?utm\_source=chatgpt.com
2. Sharma, N. (2024, November 26). *What is MobileNetV2? Features, Architecture, Application and More*. Analytics Vidhya. https://[www.analyticsvidhya.com/blog/2023/12/what-is-](http://www.analyticsvidhya.com/blog/2023/12/what-is-) mobilenetv2/#h-what-is-mobilenetv2
3. *Tomato leaf disease classification by combining EfficientNeTV2 and a SWIN transformer*. (n.d.). ResearchGate. https://[www.researchgate.net/publication/383452058\_Tomato\_Leaf\_D](http://www.researchgate.net/publication/383452058_Tomato_Leaf_D) isease\_Classification\_by\_Combining\_EfficientNetv2\_and\_a\_Swin\_Tr ansformer?utm\_source=chatgpt.com