GRA41574 (Big) Data Curation, Pipelines and Management

Background

Predicting diseases based on symptoms is one of the fundamentals of data science applications in modern healthcare. It allows early-stage diagnosis, efficient resource allocation, and better patient health outcomes. In recent years, machine learning (ML) and deep learning techniques have proven to be powerful tools in predicting and diagnosis of disease, using medical data to identify patterns and correlations that human intuition might overlook (Rastogi et al., 2022). Previous works relating to this area of research show current most focus on building ML models for diagnosing specific and concentrated diseases or rare diseases, which are considered highly lethal, urgent, and solutions should be delivered in a limited time (Ahsan et al., 2022), with the focus mainly on high-income countries with available resources.

This project explores how simplified ML models can accurately predict common diseases based on symptoms in a dataset reflecting typical diagnostic scenarios. Unlike prior studies focused on rare or acute diseases in high-resource settings, this work aims to create scalable, affordable solutions for diagnosing illnesses in underserved regions. The dataset contains symptoms and corresponding diagnosed diseases, which makes ML models typically ideal for learning patterns and making accurate predictions. Potential implications of this research are to enhance healthcare providers' decision-making and improve accessibility to medical expertise in underserved regions, where healthcare professionals are scarce for overwhelmed populations. Such models can support healthcare workers in making informed decisions under resource constraints, as it is difficult to provide proper diagnostic procedures in conditions where medical checks and treatments can be expensive and inaccessible for low-income groups (Rahman et al., 2023). Low-income countries, communities and remote areas have high barriers to accessing healthcare services, also most likely exposed to a higher risk of illnesses (Health, Income, & Poverty, 2018), posing the main contribution to worsen their health outcomes. Therefore, accurate disease prediction models could assist healthcare professionals in precise highly, time-saving diagnostics, and providing preliminary diagnoses in remote consultations. Implementing the idea into ML models would not only lessen a great burden in those countries and areas but also successfully make a breakthrough in the industry and have great impacts. Moreover, these models could encourage proactive healthcare behaviors from patients, particularly when it is easily accessible.

The goal of this report is to contribute to lifting the standard of services in the healthcare system, focusing on providing services in rural areas and low-income regions. By assessing the performance of ML models in disease prediction, this study aims to identify gaps, limitations, and opportunities for improvement. This will serve as industrial support for developing affordable, reliable, and ethical AI solutions that can transform healthcare delivery and achieve more balanced global health outcomes.

ML techniques are appropriate for this dataset for some reasons. It can handle data complexity when the dataset contains non-linear relationships and complex interactions between symptoms and diseases (Liang et al., 2022), making it challenging for traditional statistical methods to achieve accurate predictions. ML methods have been widely utilized in the healthcare industry in terms of improving disease diagnosis and precision of medicines (Oh et al., 2019). These two traits highlight ML’s strength over human

diagnosticians in terms of accuracy. Moreover, its scalability in handling large and diverse datasets efficiently, without being computationally expensive.

Methodology

The dataset used for this study was extracted from Kaggle and contains 4,920 observations with 17 features. All features represent categorical data in *object* type and represent symptoms according to outcome diseases. Observing the dataset structure, the first three symptoms are consistently present in all entries, suggesting that a minimum of three symptoms is required for diagnosis. However, accurate diagnosis often requires more than three symptoms due to overlaps in symptoms across multiple diseases. Some diseases show a broad range of symptoms, while others do not. The greater the combination of symptoms, the less likely it is that the diagnosed disease will occur.

The dataset includes 41 diseases among nearly 5,000 observations, with the number of symptoms associated with each disease varying between 3 and 17. The most frequently observed disease in the dataset is "*Fungal infection*". The most common symptoms, based on their frequency, include "*vomiting*", "*fatigue*", "*high_fever*", and "*nausea*".

The dataset is well-balanced, with each disease observed approximately 120 times. However, the diseases tend to have several root causes and not a great variety of diseases are recorded within this dataset since it is simulated and designed for educational and research purposes. These may limit its completeness and ability to capture the complexity of data in real life. Duplicate rows were not removed because each observation represents a unique diagnosed case, even if multiple patients have the same disease and identical symptoms. This approach preserves the integrity of the dataset as it reflects real-world occurrences of diseases and their symptoms. However, the dataset is missing demographic data, such as patient age and region, which could enhance this study's applicability.

Data Visualization & Pre-Processing

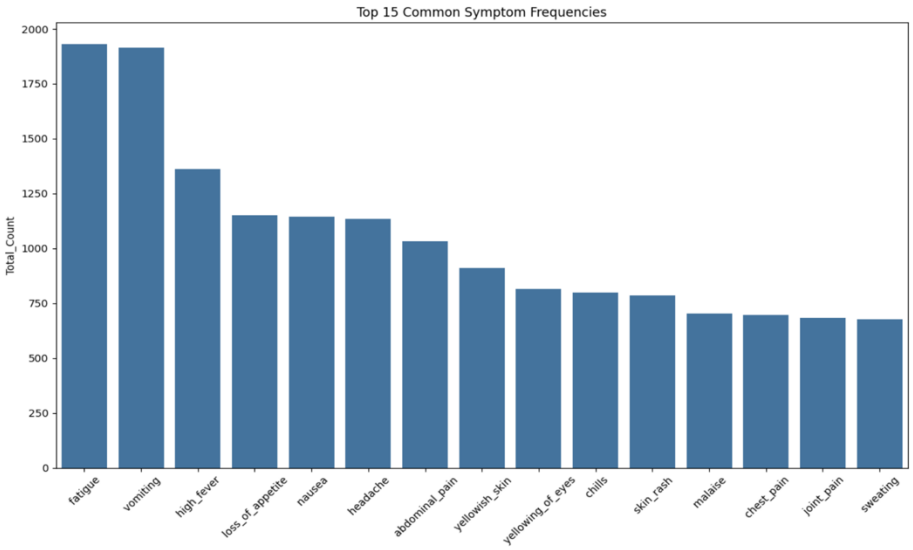

Figure 1: Top 15 Common symptom Frequencies

Figure 1 shows the frequency of the top 15 most common symptoms in the dataset. Symptoms like 'fatigue' and 'vomiting' appear most frequently, indicating their broad relevance across multiple diseases.

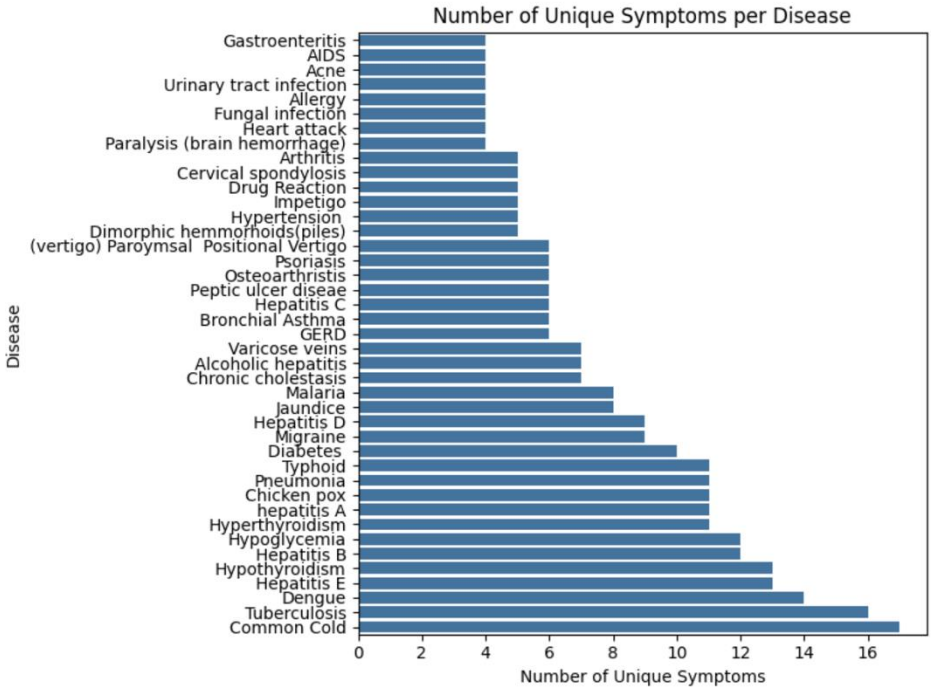

Figure 2: Number of Unique symptoms

Figure 2 demonstrates the variability in the number of unique symptoms per disease. Diseases like '*Tuberculosis*' and '*Common Cold*' are associated with up to 17 symptoms, whereas others, such as *'Acne*' have fewer, reflecting the diversity in diagnostic complexity. Some missing values indicate that there are no additional symptoms and not something that needs to be filled out. The dataset contains over 130 unique symptoms, reflecting the diversity and complexity of the features.

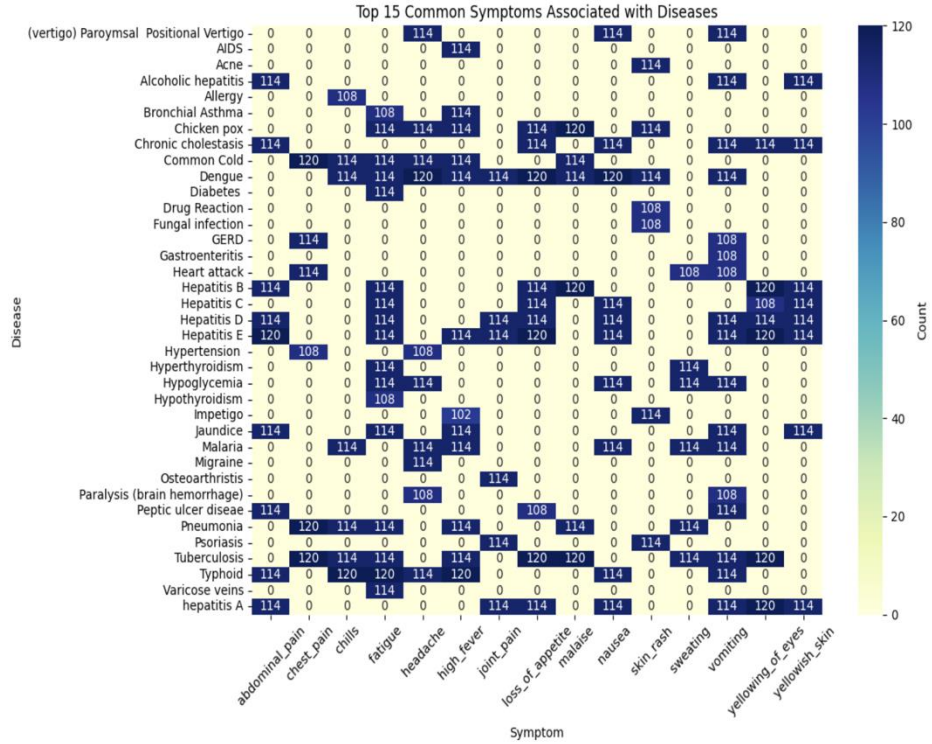

Figure 3: Top 15 Common Symptoms Associated with Diseases

Figure 3 illustrates the number of times a symptom is associated with a disease. For example, *'fatigue'* and *'vomiting'* are linked to the highest number of diseases, highlighting the complexity of diagnosing diseases based on overlapping symptoms. For more insightful visualizations, refer to the main.py file in our repository.

Predicting the model

To streamline machine learning, the dataset has been transformed into a binary format, where all symptoms were used as features and have the values 1 or 0 to indicate if a symptom occurs in the diagnosis of that disease. This transformation not only reduced the complexity of the dataset but also made it more compatible with tree-based algorithms like Decision Tree, Random Forest, and Gradient Boosting, which are known to perform efficiently with binary input features. Because of the overfitting problem, features

with high pairwise correlations were examined, and one feature from each highly correlated pair was removed to improve generalizability and reduce redundancy. The dataset is split into 70% for training and 30% for testing to achieve a balanced and unbiased evaluation. These models and their results are detailed in the Findings section.

Results

***Machine Learning Models*:** Given the multi-class nature of this problem, selecting machine learning algorithms capable of effectively handling such tasks is crucial. Three chosen algorithms have proven their effectiveness for multi-class classification tasks, with different capabilities of handling categorical outcomes effectively. To get the best possible result, Decision Tree (DT), RF (RF) and Gradient Boosting (GB) were tested to see which returned the best predictive performance. To streamline the ensemble models, features with high correlations were removed to reduce redundancy and improve predictive power in the RF and GB models' architecture.

Decision Tree: serves as the baseline model and the simplest among the three. This model creates a tree structure by splitting features recursively. In each level a condition is presented, selecting the feature that best splits the data based on the selected criterion (*Decision Trees | Machine Learning*, n.d.). The other two models are based on DTs but implement multiple trees differently and use unique algorithms to reach their results.

Random Forest: gathers multiple DTs and finds the average result of multiple predictions from different trees. This increases the robustness and generalizability of the model, and reduces the impact of individual tree biases, leading to more stable and accurate predictions by aggregating the predictions from multiple DTs (*Random Forest | Machine Learning*, n.d.).

Gradient Boosting: is also an ensemble of DTs, but instead of finding the average of the predictions, it runs multiple trees and tries to learn from the mistakes of the previous model further developing the model's predictive power and learning complex patterns from the data (*Gradient Boosted Decision Trees | Machine Learning*, n.d.).

The models were implemented using *scikit-learn*, where the hyperparameters were optimized using Grid Search for DT and GB. For instance, a maximum depth of 7 was limited to the GB, and for the learning rate 0.05. Meanwhile, RF was fine-tuned by removing one feature of a highly correlated pair of features. The GB also used these reduced features as its X variable. The classification report from the *scikit* library was used to display metrics such as *accuracy*, *precision*, *recall* and *F1-score*, providing a good overview of the predictive power for each disease. Metrics such as precision and recall are critical to ensure balanced performance across all the different diseases.

Model	Accuracy	Precision	Recall	F1-score
Decision Tree	0.98	0.99	0.98	0.99
Random Forest	0.96	0.97	0.96	0.96
Gradient Boosting	0.97	0.98	0.97	0.97

Table 1: Metrics results are gathered from the classification report printed out from the main.py

Findings

Table 1 summarizes the performance metrics for the three models on the test dataset. The DT achieved the highest metrics across all categories.

However, these results likely indicate overfitting as this kind of model tends to learn the training data too well and might not generalize with unseen data. On the other hand, the RF model showed slightly lower

results, but this can show its ability to maintain robust performances across the different diseases. GB achieved high accuracy and balanced cross-validation performance, suggesting improved generalization compared to DT.

Learning Curves

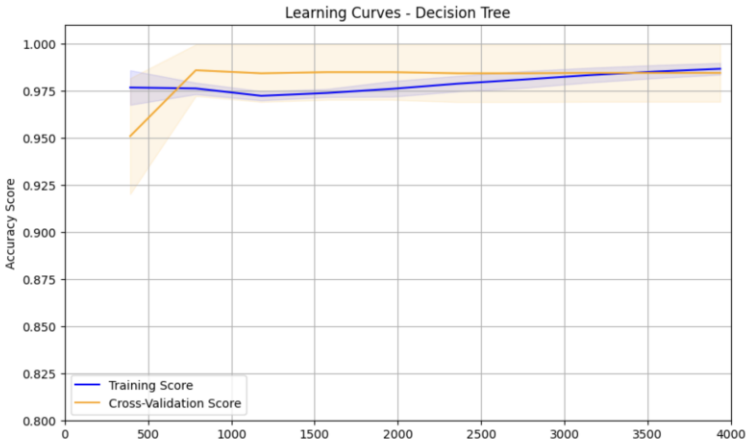

Figure 4: Learning curve for DT

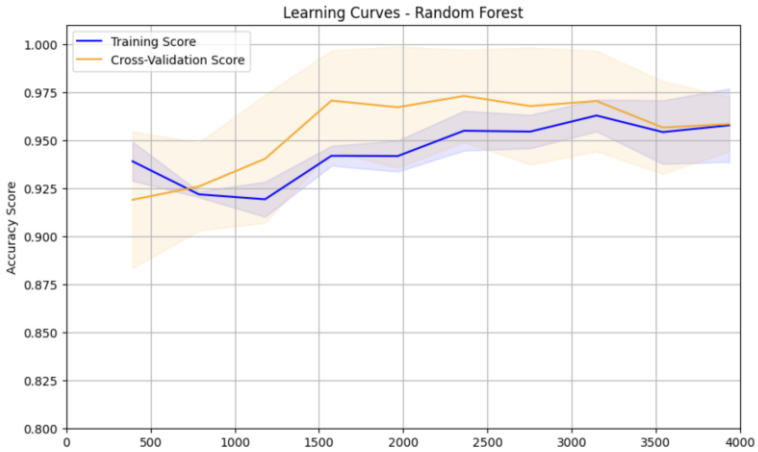

Figure 5: Learning curve for RF

To further investigate the models' behavior, the average learning curves for the training and cross-validation sets are visualized through graphs. These curves show how well the model performed when the dataset size increases. DT and GB learning curves display a similar pattern, while RF converged well but highly fluctuated. DT (*Figure 4*) has a high accuracy score in both training and cross-validation.

However, examining it closely shows that its cross-validation curve started to fall under the training at the end, showing that the model has a slim chance of overfitting. This indicates a limited ability to generalize beyond the training data of this model. RF (*Figure 5*) has a lower accuracy score but converges at 0.96. This proves a well-balanced model and predicts unseen data, avoiding overfitting as one of its strengths. Finally, GB (*Figure 6*), achieved steady improvements in both training and cross-validation. It has a smooth progression as the training set increases and converges at around 97.5% displaying its power to refine predictions while minimizing overfitting. As mentioned, the DT has a similar progression, however, after the point of convergence, the DT's curves got separated. This indicates that GB improved the potential of overfitting better.

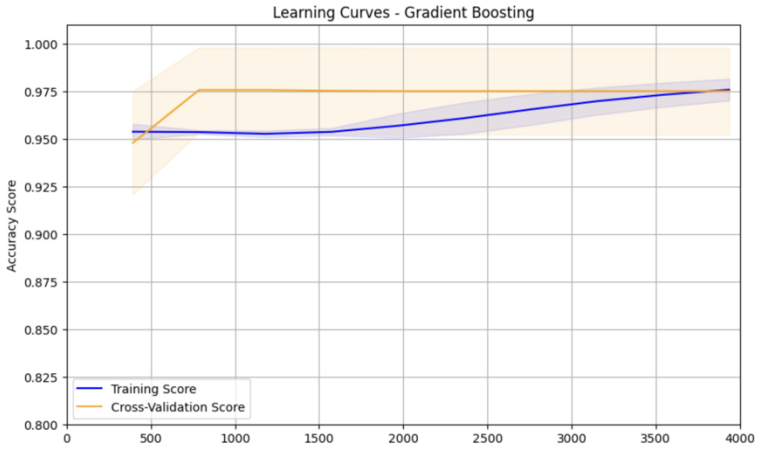

Figure 6: Learning curve for GB

Evaluation

To evaluate how well each model classified the diseases, the confusion matrices were plotted using a heatmap for better visualization (*Figure 7*). As all models do not achieve 100% accuracy, meaning that some diseases have been misclassified. For DT, the most common disease that gets misclassified is "*drug*

reaction". This pattern suggests that "*drug reaction*" may share symptoms with other diseases, which the DT struggled to differentiate. The other two models misclassified the diseases with *"diabetes"*, even after high correlation pairs were removed, explaining that *"diabetes"* also has some similar symptoms to other diseases. (*Refer main.py of CM for RF & GB*) [1]

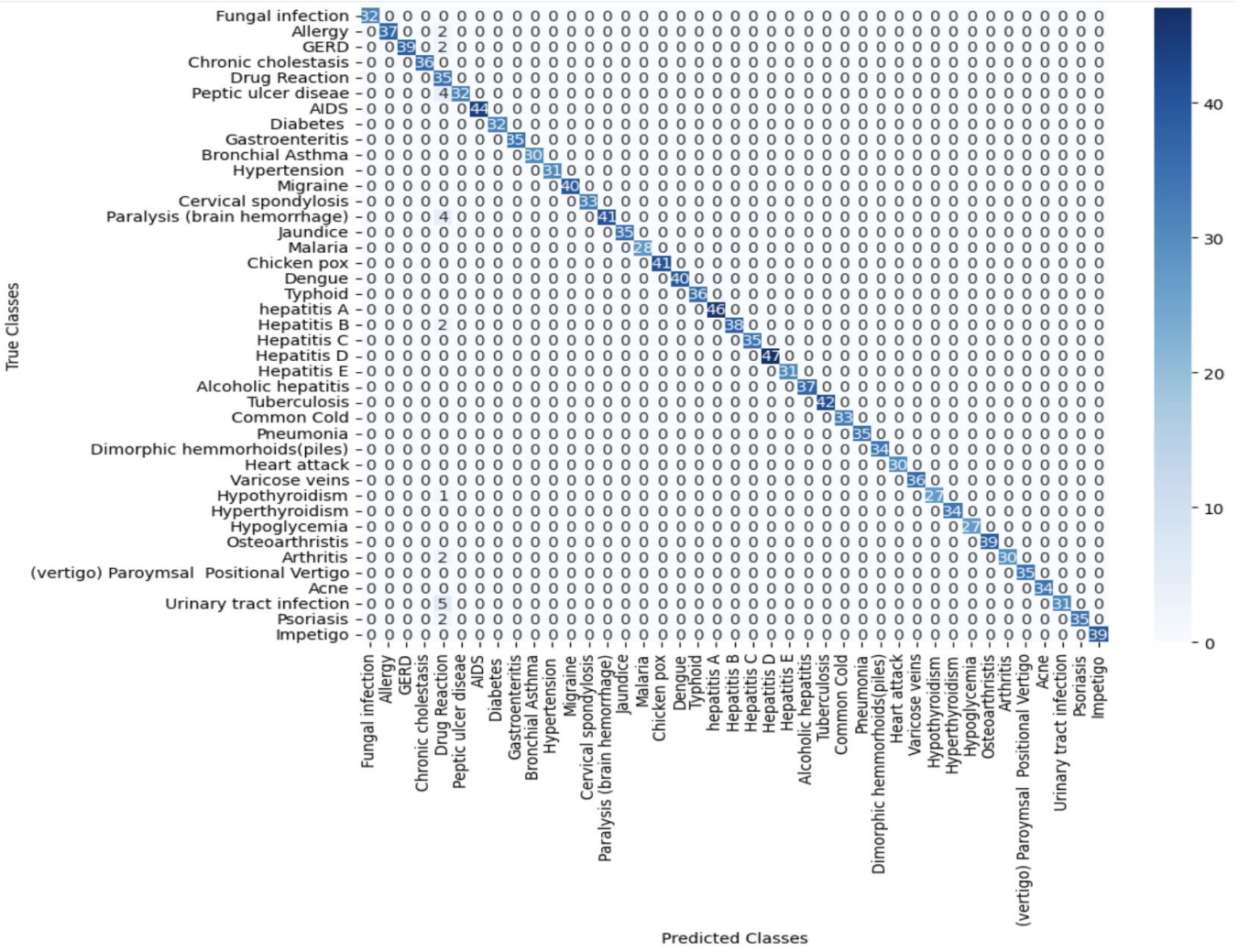

Figure 7: Confusion Matrix by Decision Trees result

Conclusion

Key findings

DT achieved competitive results, with *accuracy* and *F1-scores* metrics comparable to RF and GB. Its strong performance can be used for the dataset's simple structure, particularly the presence of duplicate rows, simulating different patients' records with the same diagnosed diseases, for the model to learn specific patterns. However, this reliance on consistency could limit its generalization to more diverse or expanded datasets.

[1] Due to the limited pages, please refer to (main.py Line 97 and 129) for confusion matrices of RF & GB results.

RF demonstrated a robust prediction, converging with around 96% accuracy. This close alignment confirms that the model avoids overfitting while maintaining a high level of accuracy across the dataset. While this model performs well, it has some reliance on ensemble methods which could sometimes lead to computational inefficiencies.

GB converges around 97,5% accuracy, only a minimal difference from DT's. This is quite good as it clearly shows this model's ability to represent the learning parameters well, and understand complex patterns, which makes it suitable for the task.

In summary, all three models showed great performance for multi-class classification, with GB emerging as the most robust. However, given the simplicity of the dataset, the DT also delivered strong results and gave a little less misclassification of the GB, making it a viable option for straightforward tasks, while RF offered a balanced alternative.

Practical Relevance

The ability of GB to classify diseases accurately based on symptoms indicates how it can be used to improve diagnostic processes in healthcare, particularly useful in scenarios where medical resources are limited. Automation helps in assessing the patient faster although it must be used cautiously and someone with a medical background to confirm the model's diagnosis. Using GB in this case could pose some challenges in scaling its application to real-time diagnostics, but in combination with good software, we can streamline this to work faster.

Limitations & Future Directions

One major limitation of this study is the nature of the dataset. Even when balanced, it does not fully represent the complexity of real-world medical data, as they often come with many more features, for instance, the patient's medical profile & history, and demographic details. The presence of the duplicated rows, though justified for this context, may oversimplify certain patterns. Furthermore, the focus on symptom-based classification alone excludes other important factors that could influence the final diagnostics.

Another key limitation of this study is the lack of direct input from healthcare professionals or domain experts during feature selection and model evaluation. While the models demonstrated strong performance, feature importance can pose the risk of removing critical features or introducing biases, that may compromise the model's performance. Those removed features might appear irrelevant to a machine learning perspective yet could carry significance in a clinical setting. Therefore, it is essential to collaborate with healthcare professionals to ensure everything is aligned properly.

Future implementations could include integrating these models into electronic health record (EHR) systems to streamline diagnostic processes and improve clinical decision-making. Developing mobile applications for symptom checkers could empower individuals to perform preliminary self-assessments, especially in regions with limited access to healthcare professionals. People can take some precautions, for example, improving their intake and eating habits, carrying out exercises, etc. while waiting for the doctor's appointment. Additionally, these tools could enable more personalized data-driven healthcare practices by incorporating specific patient data.

Expanding the dataset to include diverse populations and additional features, such as lab results or treatment outcomes, could enhance the generalizability of the models. Further research could also focus on refining hyperparameters or exploring more advanced algorithms such as neural networks. Lastly, improving the model's interpretability and addressing ethical concerns, such as biases in predictions, will be crucial to implying data science applications constructively and trustfully to the healthcare industry.

GitHub repository link: https://github.com/S1098612/GRA4157-Final-Project

Reference

Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022). Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. *Healthcare*, *10*(3), 541. https://doi.org/10.3390/healthcare10030541

Decision trees | Machine Learning. (n.d.). Google for Developers. Retrieved December 9, 2024, from https://developers.google.com/machine-learning/decision-forests/decision-trees

Gradient Boosted Decision Trees | Machine Learning. (n.d.). Google for Developers. Retrieved December 9, 2024, from https://developers.google.com/machine-learning/decision-forests/intro-to-gbdt

Health, Income, & Poverty: Where We Are & What Could Help. (2018). Project HOPE. https://doi.org/10.1377/hpb20180817.901935

Liang, D., Frederick, D. A., Lledo, E. E., Rosenfield, N., Berardi, V., Linstead, E., & Maoz, U. (2022). Examining the utility of nonlinear machine learning approaches versus linear regression for predicting body image outcomes: The U.S. Body Project I. *Body Image*, *41*, 32–45. https://doi.org/10.1016/j.bodyim.2022.01.013

Oh, J., Yun, K., Maoz, U., Kim, T.-S., & Chae, J.-H. (2019). Identifying depression in the National Health and Nutrition Examination Survey data using a deep learning algorithm. *Journal of Affective Disorders*, *257*, 623–631. https://doi.org/10.1016/j.jad.2019.06.034

Rahman, S. M. A., Ibtisum, S., Bazgir, E., & Barai, T. (2023). The Significance of Machine Learning in Clinical Disease Diagnosis: A Review. *International Journal of Computer Applications*, *185*(36), 10–17. https://doi.org/10.5120/ijca2023923147

Random Forest | Machine Learning. (n.d.). Google for Developers. Retrieved December 9, 2024, from https://developers.google.com/machine-learning/decision-forests/intro-to-decision-forests

Rastogi, M., Vijarania, D. M., & Goel, D. N. (2022). *Role of Machine Learning in Healthcare Sector* (SSRN Scholarly Paper 4195384). Social Science Research Network. https://doi.org/10.2139/ssrn.4195384