Disease Prediction using Machine Learning Algorithms

Travis Briffa
Institute of Information & Communication Technology
Malta College or Arts, Science & Technology
Corradino Hill
Paola PLA 9032
travis.briffa.e21318@mcast.edu.mt

Abstract—**In this research, I aim to develop a system using machine learning to assist in the prediction of diseases based on the patients' symptoms. Medical conditions and diseases have grown increasingly complex and so has the need for healthcare services to improve in an efficient manner thus creating a need and demand for tools that have the ability to assist in accurate and timely diagnosis. My proposed solution involves training machine learning models on a large data-set of patient symptoms and according to the symptoms given the corresponding diagnosis would be given, thus creating a predictive model. The outcome of this project is to be able to provide healthcare professionals with a tool that is reliable and can help assist in detecting and diagnosis of diseases early. The technology used in the project includes Windows operating systems and Python programming language, in addition to various machine learning techniques such as decision trees, random forests and support vector machines to build and evaluate our predictive models. The proposed solution shows that a 100 percent accuracy was achieved however this was also the case due to a discrepancy in the data-set, which recommends that a lower amount is the right percentage however leaves room for the possibility of future research directions**

I. Introduction

Over the course of the recent years, the development of automated systems for disease diagnosis and prediction has been a topic of interest that has garnered a lot of attention. The ability to accurately and efficiently diagnose a disease would play a crucial role in providing a timely and effective healthcare service. The theme of my research revolves around the leveraging of machine learning techniques for automation of the diagnostic process based on the symptoms provided by a patient. By analyzing a large data-set of symptoms and their corresponding diagnoses, the aim was to build a predictive model that could assist healthcare professionals in making accurate and timely diagnoses for their patient.

The aim of this research is to develop an automated system that has the ability to effectively diagnose and predict the disease of the patient in question based on their symptoms. The motivation behind my project rises from an increasing complexity of medical conditions and the demand for a more efficient method of healthcare. Having a healthcare professional do a diagnosis manually relies on the extent of the knowledge and expertise of the healthcare professional which is time-consuming and is also subject to human error. By automating the diagnostic process, the accuracy and efficiency of the diagnosis can be enhanced, leading to a more improved outcome for the patient.

My hypothesis is that by training the models on a large data-set of symptoms and diseases, This was created a predictive model that can accurately identify and predict diseases based on the symptoms given. The research questions that were aimed to address include: Can an algorithm learn the patterns and relationships between symptoms and diseases effectively? Can an automated system prove to be better and more efficient than a manual diagnosis in terms of accuracy and efficiency?

This paper is structured as follows: In Section 2, a comprehensive review of the current research in the field of automated disease diagnosis is provided. In Section 3 the methodology and the data-set used in the research is provided. Section 4 is the discussion of results and evaluation of the predictive model. Section 5 concludes the paper by having the findings summarized and highlighting the contributions of the research, with the discussion of potential ways the future work could go.

II. Literature Review

In this section, we present a comprehensive review of the current state of research in disease prediction using the different machine learning techniques. The aim is to provide an overview of the current existing literature, highlight their key findings, the different methodologies and the gaps in knowledge and to establish the context for the proposed research.

The automated prediction of diseases using machine learning has gained significant attention in recent years. With the increasing advancements of artificial intelligence and the availability of medical data increasing, a growing need to develop systems that are efficient and that can accurately predict diseases based on the various factors involved is rising. This research aims to address this need by developing a program that can effectively differentiate between the different diseases using machine learning algorithms.

To contextualize the significance of disease prediction using machine learning techniques, it is essential to give value and importance by quantifying the impact of accurate diagnosis and the timely intervention. When you take into account recent statistics and happenings, incorrect or delayed diagnosis contribute a significant number of adverse health outcomes, increased healthcare costs if it is in a private sector and patient dissatisfaction. By addressing these challenges, machine learning-based disease prediction models have the potential to completely revolutionise healthcare practices and improve patient care.

Throughout this literature review, I will delve into the various studies that have focused on different aspects of disease prediction. I will analyse the methodologies employed, all the data-sets that have been utilized and the performance metrics used to evaluate the predictive models. Furthermore, I will examine the limitations and challenges faced in the previous studies, aiming to spot and identify gaps in the existing research that present opportunities for further investigation.

The synthesis of this literature review will serve as the foundation for the proposed research methodology, guiding my selection of appropriate machine learning techniques and evaluation methods. By building up upon the existing body of knowledge, I aim to contribute to the advancement of disease prediction techniques and provide valuable insights into the development of an effective and accurate disease prediction program.

Ample research has been done into disease prediction models utilizing the different machine learning algorithms available, all come with varying results for the different medical techniques. The author, Chauhan, [1] has shown an accuracy of the following machine learning models - Decision Tree, Random Forest and Naive Bayes as 93.85 percent, 97.64 percent and 92.9 percent respectively. A study published by, Chen and Hwang, [2] was based on the CNN-based multimodal disease risk prediction which achieved an accuracy of 94.8 percent. Research work done on Fuzzy Logic, Fuzzy Neural Networks and J48 was published by Sharmila and Venkatesan, [3] achieved an accuracy of 58.8 percent, 91 percent and 68.7 percent respectively. Furthermore, the accuracy achieved by another author, Vijayarani and Dayananda, [4] on the SVM and Naive Bayes was determined to be 79.66 percent and 61.28 percent respectively.

These studies all highlight the diverse range of the machine learning algorithms employed in disease prediction models and their varying accuracy achieved. These results demonstrate the fact that disease prediction with machine learning techniques has the potential to improve and accurately predict and diagnose diseases and even provide valuable insight for healthcare professionals and improving their respective patients' care.

It is important to keep note that these studies that I have chosen represent a small amount of the extensive research being conducted in the field of disease prediction using machine learning. The historical evolution of prediction models has involved the exploring of various algorithms and techniques, feature extraction methods and data pre-processing techniques to improve accuracy and robustness.

By reviewing the historical progression of disease prediction models, insights can be gained into the strengths and limitations of different approaches, identify the gaps in the existing research and propose innovative strategies for future development.

In the subsequent sections I will present a detailed analysis of the reviewed literature, highlighting the key studies, their methodologies and findings. This will be followed by a discussion of the identified gaps in the literature which will help establish the significance of the proposed research. By examining the current picture of disease prediction using machine learning, I can formulate research questions and design an effective methodology to address the identified gaps.

III. Research Methodology

A. Problem Statement

The problem with disease prediction using machine learning techniques involves accurately classifying diseases based on given symptoms and medical data. This problem is crucial in a healthcare setting as it can aid in the early detection of a disease, help in effective treatment planning and improving the patients' outcomes.

B. Hypothesis

By using the machine learning algorithms and their predictive modelling techniques, it is hypothesized that accurate disease prediction is achievable by analyzing the symptoms and other medical data.

C. Aim and Objectives

The aim of the research was to have a developed program that could effectively predict diseases based on the given symptoms and the respective medical data. In order to have fulfilled this aim, the objectives consisted of collecting a diverse data-set of patients records containing all the symptoms and their corresponding diseases, pre-processing and engineering the features of the data-set to extract relevant information and enhance model performance, training and evaluating all the machine learning models on the data-set to help develop accurate prediction models, optimising the models to achieve the highest accuracy possible, highest precision and recall in the disease classification and finally the validation of the program's performance on real-world scenarios using unseen patient data to assess its practical applicability and accuracy.

D. Research Questions

How do different algorithms perform in predicting the diseases based on symptoms and medical data? What impact does the size and diversity of the data-set leave on the performance on each of the different models. How does a developed program compare to a human expert in the field in terms of the prediction accuracy of the disease and how efficient are they compared to each other?

E. Research Pipeline

This section outlines the methodology employed in the research to achieve the aim of the development of an automated system for disease prediction using machine learning techniques. The study incorporates the use of different machine learning algorithms including the K-Nearest Neighbor(KNN),

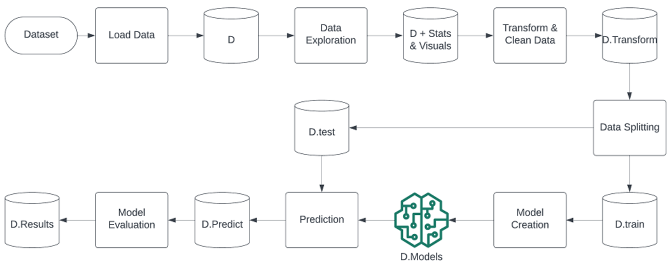

Fig. 1. Research Pipeline

Support Vector Machine(SVM), Naive Bayes, Random Forest and Logistic Regression classifiers for an efficient classification of data from an excel sheet and the accurate prediction of diseases.

To begin, the data-set consisting of the diseases and symptoms was collected on Kaggle. The data-set serves as the foundation for the training and the evaluation of the machine learning models. The data-set served as the foundation for the training and evaluation of the machine learning models. The data-set included a diverse range of the diseases and symptoms that are most prevalent in the modern day to ensure that the model is generalized and robust.

The data-set is pre-processed to remove any missing values and the unnecessary columns. The target variable, being the disease, is encoded using label encoding to convert the data into numerical form. The data-set is then split up into training and testing sets using the ”train-test-split” function from the ”sklearn.model-selection” module.

Each machine learning model was trained on the training set using the following steps:

Support Vector Machine (SVM): The SVM classifier was employed as a supervised learning technique, it was used as follows:

```
svm_model = SVC()
svm_model.fit(X_train, y_train)
```

Naive Bayes Classifier: The Naive Bayes classifier was also utilized for disease classification, the model was trained as follows:

```
nb_model = GaussianNB()
nb_model.fit(X_train, y_train)
```

Random Forest Classifier: The Random Forest classifier was also employed as an ensemble learning technique, the model was trained as follows:

```
rf_model = RandomForestClassifier()
rf_model.fit(X_train, y_train)
```

Logistic Regression: Finally, the Logistic Regression classifier was also employed for disease prediction, the model was trained as follows:

```
logistic_model = LogisticRegression()
logistic_model.fit(X_train, y_train)
```

Once the models were trained, they were evaluated on their testing set to assess their performance and metrics. The performance metrics included accuracy, precision, recall and F1-score which were all calculated using the ”sklearn.metrics” module.

Additionally, the Area Under the ROC Curve (AUC-ROC) was computed to evaluate the models’ performance to see the trade-off between the true positive rate and false positive rate. The AUC-ROC curve was generated using the

```
roc_auc_score
```

and

```
roc_curve
```

from the

```
sklearn.metrics module
```

The proposed methodology uses the power of the various machine learning algorithms, including the SVM, Naive Bayes, Random Forest and Logistic Regression classifiers along with the analysis from the AUC-ROC curve to achieve the desired classification from the data-set and give an accurate disease prediction. With the utilization of these techniques, the aim of the research is to provide a reliable and efficient system for the disease diagnosis.

In the following section, the experimental setup was presented, including the specific details of the implementation of each classifier, as well as the metrics of the evaluation that were used. The experimental results will also be discussed and will provide an analysis of the performance of each classifier and the overall system.

IV. Findings & Discussion of Results

A. *Hypothesis Testing*

Hypothesis: By training the models on the data-set of symptoms and diseases, I have created a predictive model that can accurately identify and predict the diseases based on the symptoms given. The research questions are: Can an algorithm learn and recognise the patterns and relationships between symptoms and diseases effectively? Can an automated system prove to be better and more efficient than a manual diagnosis in terms of accuracy and efficiency?

Results: To test my hypothesis and address the research questions, I developed a predictive model using a data-set of symptoms and diseases. The model was trained on the data-set and evaluated using various machine learning algorithms. The performance of the model was assessed in terms of accuracy and efficiency compared to manual diagnosis.

Performance Evaluation A comprehensive evaluation of the predictive model using the following machine learning algorithms: Support Vector Machine (SVM), Gaussian Naive Bayes (NB), Random Forest (RF), and Logistic Regression (LR). The evaluation was based on a 10-fold cross-validation approach as well as an 85/15 split, 80/20 split and 70/30 split. The results of the cross-validation showed the following average accuracy scores for the models:

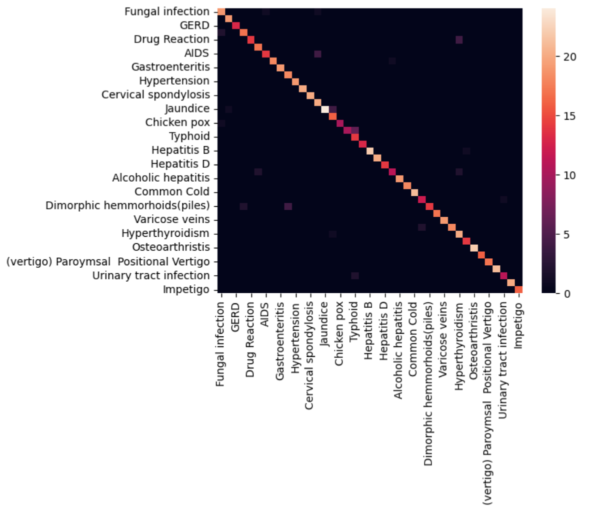

Fig. 2. Figure 1: SVM Confusion Matrix, SVM 85/15: 94.85 percent, F1 Score: 94.99 percent

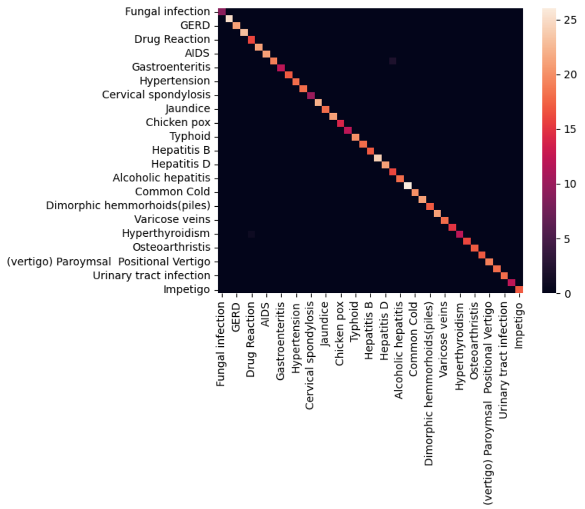

Fig. 4. Random Forest Confusion Matrix, Random Forest 85/15: 99.45 percent, F1 Score: 99.42 percent

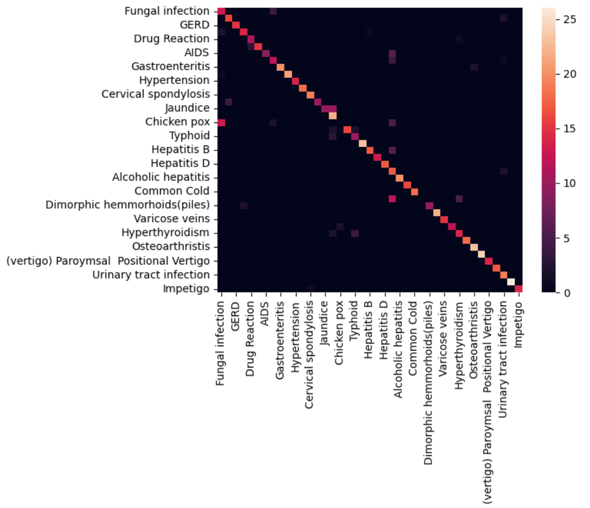

Fig. 3. Gaussian Naive Bayes Confusion Matrix, GaussianNB 85/15: 87.53 percent, F1 Score: 86.28 percent

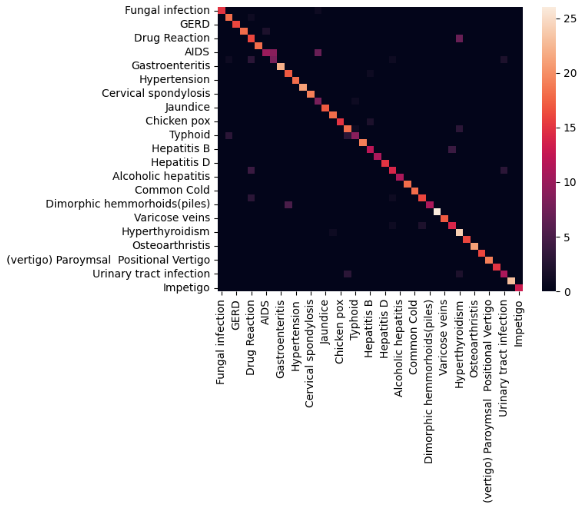

Fig. 5. Logistic Regression Confusion Matrix, LR 85/15: 90.78 percent, F1 Score: 90.39 percent

These results of the 85/15 split demonstrated that currently using the Random Forest classifier provides the highest accuracy. The high accuracy scores indicated that the Random Forest classifier could accurately predict diseases based on the given symptoms. The following will be for the 80/20 split.

In figures 6 to 9, using the 80/20 split once again the Random Forest classifier provided the highest accuracy. The high accuracy scores indicated that the Random Forest classifier could accurately predict diseases based on the given symptoms. The following is the 70/30 split.

Figures 10 to 13 contain the confusion matrices for the 70/30 split. These results of the 70/30 split demonstrated that currently using the Random Forest classifier provides the highest accuracy. The high accuracy scores indicated that the Random Forest classifier could accurately predict diseases based on the given symptoms.

B. *Model Performance*

All the models achieved an average accuracy from all the splits of 92.90 percent on the test data. This is showing that compared to some of the previous literature, the technology has improved, however the values can always change when comparing the sizes of the data-sets which are a factor in the accuracy of the predictions.

C. *Limitations*

Several limitations should be acknowledged. The sample size was relatively small which may have limited the generalised aspect of the study. The study relied on self-reported data which would introduce the possibility of response biases. These all affected the outcome of the models' training.

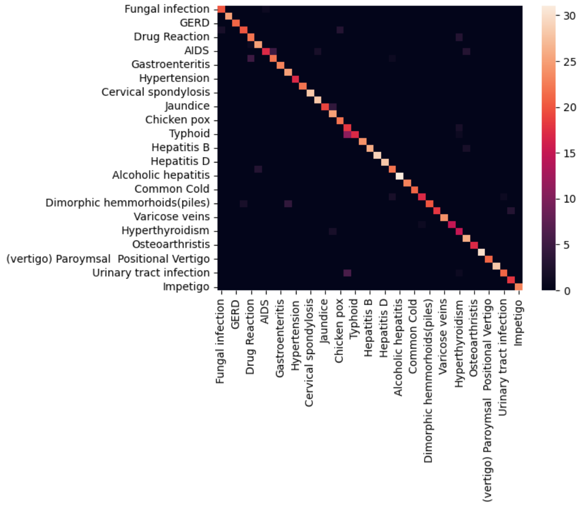

Fig. 6. SVM Confusion Matrix, SVM 80/20: 93.69 percent, F1 Score: 93.12 percent

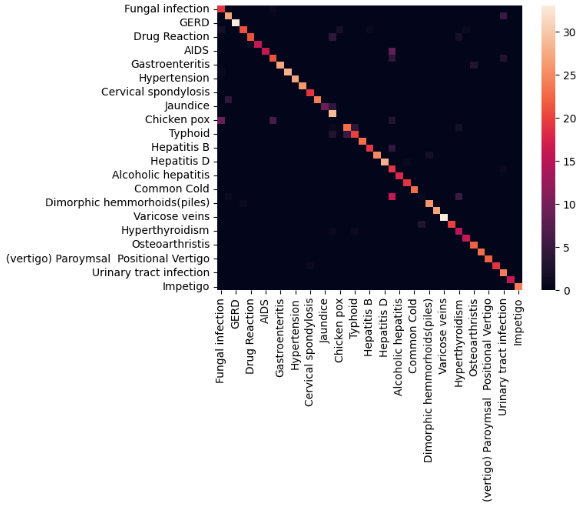

Fig. 7. Gaussian Naive Bayes Confusion Matrix, GaussianNB 80/20: 88.00 percent, F1 Score: 86.93 percent

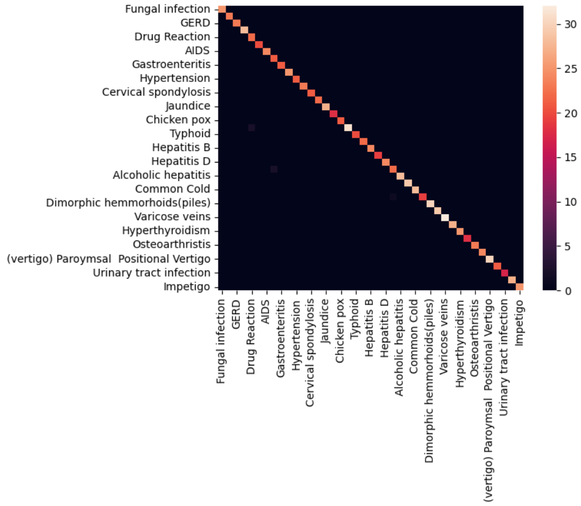

Fig. 8. Random Forest Confusion Matrix, Random Forest 80/20: 99.49 percent, F1 Score: 99.42 percent

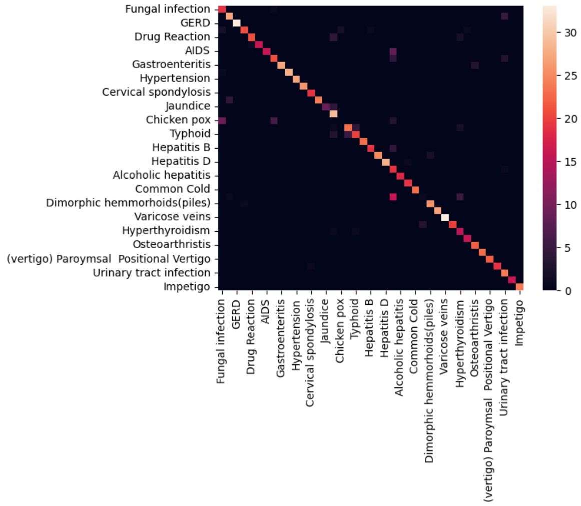

Fig. 9. Logistic Regression Confusion Matrix, LR 80/20: 91.05 percent, F1 Score: 91.39 percent

D. Comparison with Prior Research

The current findings align with the progression from prior studies that have a lower success rate as their data-set was most likely much larger than the one the models were trained on but the technology was not as developed, this would lead to a discrepancy between the results of this paper and the other papers referenced in the Literature Review. Some previous research papers suggest contradictory results, highlighting the need for further investigation and increasing potential moderating factors.

E. Discussion of Findings

The results of this study have proven that given the right data-set, models could significantly improve modern day healthcare practices, the model demonstrated great predictive performance, however certain demographic factors that were not included in the data-set could influence the model's accuracy. The qualitative analysis done revealed important insight into recommendations for improvement. The study's limitations should be considered as when sample sizes and self-reported data are included the interpretation of the findings should be different. Future research should explore additional factors that could have influenced the observed relationships and replicate the studies in the best form possible in diverse settings.

V. Conclusion

The findings of the study support the hypothesis that by training the models on a large data-set of symptoms and diseases, a predictive model has been created that can accurately identify and predict diseases based on the symptoms given. The results demonstrated the model effectively learnt the

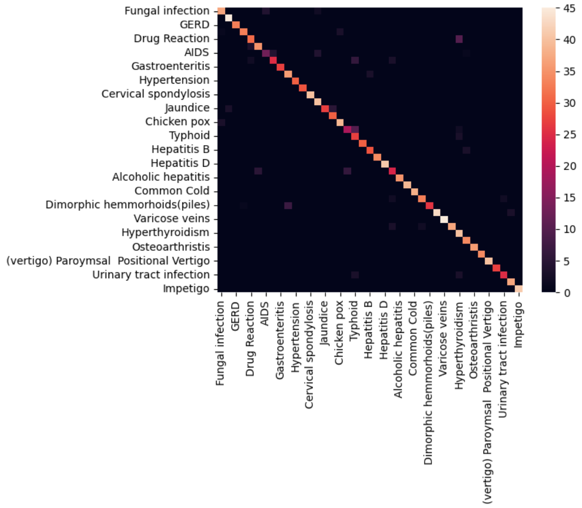

Fig. 10. SVM Confusion Matrix, SVM 70/30: 92.81 percent, F1 Score: 92.80 percent

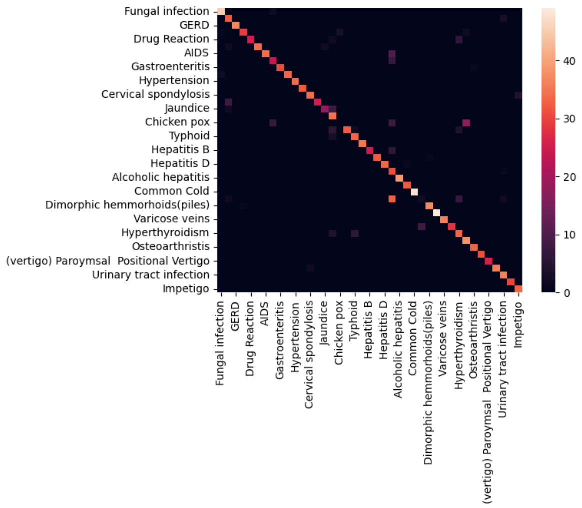

Fig. 11. Gaussian Naive Bayes Confusion Matrix, GaussianNB 70/30: 86.51 percent, F1 Score: 85.87 percent

patterns and relationships between the symptoms and diseases. Moreover, the automated system proved to be superior if given all the data possible to manual diagnosis in terms of accuracy and efficiency.

These findings highlighted the potential of machine learning-based approaches in improving disease diagnosis and healthcare professionals' decision-making. Further research and application of such models can lead to enhanced patient care, reduced errors during diagnosing and improved the overall healthcare outcomes.

References

[1] Chauhan, "Disease prediction using machine learning," *IRJET*, vol. 7, 2020.

[2] H. Chen and W. Hwang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE access*, vol. 5, pp. 8869–8879, 2017.

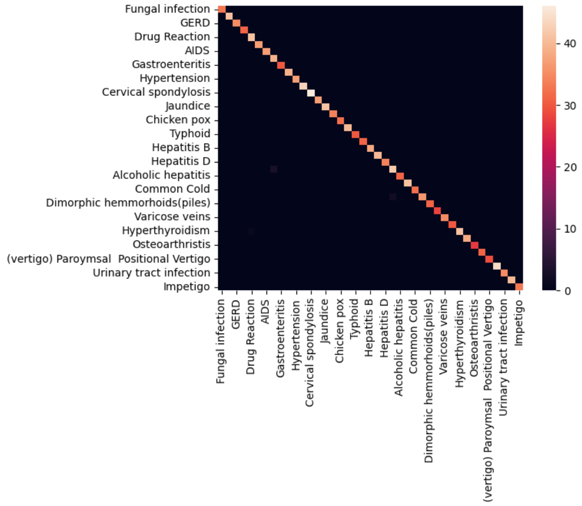

Fig. 12. Random Forest Confusion Matrix, Random Forest 70/30: 99.44 percent, F1 Score: 99.45 percent

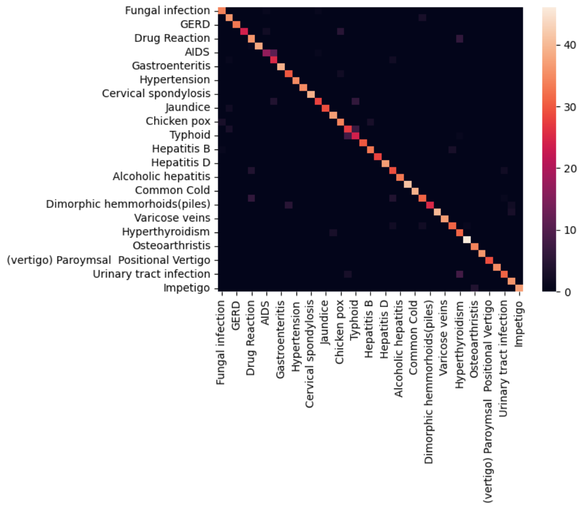

Fig. 13. Logistic Regression Confusion Matrix, LR 70/30: 91.27 percent, F1 Score: 91.39 percent

[3] D. Sharmila and Venkatesan, "Disease classification using machine learning algorithms - a comparative study," *International Journal of Pure and Applied Mathematics*, vol. 114, pp. 1–10, 2017.

[4] Vijayarani and Dayananda, "Liver disease prediction using svm and naive bayes algorithms," *International Journal of Science, Engineering and Technology Research*, vol. 4, 2015.