Disease Diagnosis Predictions from Lab Results

Shaaf Nadeem1[0000*−*1111*−*2222*−*3333] and Ibrahim Khanr2*,*3[1111*−*2222*−*3333*−*4444]

Cure MD, Lahore [muhammad.shaaf@curemd.com](mailto:muhammad.shaaf@curemd.com)

*{*abc,lncs*}*@uni-heidelberg.de

# Abstract

In the world of medical science, machine learning has become a disruptive force that offers unheard-of prospects to improve healthcare delivery, diagnosis, treat- ment, and research. An overview of machine learning’s potential and applications in the field of medical science is given in this abstract.

Utilizing computational algorithms, machine learning analyzes huge, com- plicated datasets to uncover insightful patterns, predictions, and insights. This technique has found significant application in the medical field in a number of crucial areas.

# Introduction

The findings of a machine learning (ML) effort to forecast diagnoses based on laboratory data are presented in this article. In order to find the most accu- rate predictive model, the project required the extraction of data from a SQL database, meticulous data pre-processing, and application of multiple ML tech- niques. The main goal was to use ML approaches to improve diagnosis accuracy. The data pre-processing stage was thorough and careful. It included activities like data cleaning, duplication removal, unit standardization, null value deletion, and the preservation of the most recent laboratory results. These actions played

a critical role in guaranteeing the accuracy and dependability of the dataset.

We tested with a variety of models, including Random Forest, K-Nearest Neighbors (KNN), Light GBM, Logistic Regression, and XGBoost, to assess the effectiveness of various ML algorithms. These models were evaluated based on how well they could forecast diagnoses.

This experiment, in conclusion, shows the promise of machine learning in the area of diagnostic prediction. We were able to pinpoint XGBoost as the most promising model for enhancing diagnostic accuracy by applying stringent data preprocessing procedures and analyzing other ML algorithms. The results of this study may have important ramifications for medical practitioners looking for more precise and effective diagnostic equipment.

XGBoost fared better than the other models, as seen by its F1 score of 0.36 and precision rate of 65 percent. Additionally, multilabel classification outper- formed multi-class classification, highlighting the value of simultaneously taking into account numerous diagnostic options.

# Background

## SQL Databases

**Data Structure and Relational Mode:l** The relational paradigm, which is the foundation of SQL databases, is ideal for structured data. The efficient organisation and retrieval of data through tables with rows and columns made possible by this paradigm makes it appropriate for storing structured data, such as user identifying information.

**Data Integrity:** With the help of constraints like primary keys, unique con- straints, and foreign keys, SQL databases offer ways to guarantee data integrity. By doing this, user identity information is kept accurate and reliable.

**Querying and Retrieval:** SQL databases provide a potent query language (SQL) that enables developers to carry out sophisticated queries for obtaining particular user data. This is very useful when looking for certain people or fil- tering data according to different criteria.

**Security:** Strong security features are offered by SQL databases, such as user authentication, role-based access control (RBAC), and data encryption. These functions aid in preventing unauthorised access to and breaches of user identity information.

**Backup and Recovery:** Developers can regularly create backups of user data thanks to the built-in backup and recovery capabilities found in many SQL databases. In the event of inadvertent data loss or corruption, this guarantees that data can be restored.

## Pandas Dataframes

**Data Handling and Transformation:** Developers can handle, modify, and transform tabular data effectively thanks to Pandas’ sophisticated and adaptable data structure, the data frame. Data frames are a logical solution for processing user identity data because it frequently comes in structured formats.

**Data Cleaning and Preprocessing:** Data used to identify users in the real world sometimes has inconsistent values, outliers, or missing information. For data cleaning and preparation, Pandas provides a wide range of features, such as handling missing data, getting rid of duplicates, and changing data types.

**Data Filtering and Selection:** Pandas data frames excel at conducting SQL- like joins, merges, and concatenations in applications that call for the integration of user identity data from several sources or databases. This is useful for bringing data from many sources together into a single format.

## Machine Learning

In recent years, machine learning (ML) has grown significantly in popularity as a potent computational method for drawing conclusions and making predictions from intricate and sizable datasets. The employment of ML approaches in the context of our study, which is concerned with using test results to diagnose medical disorders, offers enormous promise and significance.

**Data-Driven Medicine:** The advent of electronic health records (EHRs) and the digitization of healthcare data have ushered in an era of data-driven medicine. These vast repositories of patient information, including laboratory results, pro- vide an opportunity to harness ML to improve patient care, streamline diagnosis, and enhance healthcare decision-making.

**Challenges in Medical Diagnosis:** Medical diagnosis is often intricate, influ- enced by numerous factors, and sometimes requires the consideration of a wide range of clinical parameters. Additionally, diseases can manifest differently in each patient, making accurate diagnosis a complex task. ML models, with their ability to uncover patterns and relationships in data, are well-suited to address these challenges.

**Feature Engineering:** In our project, the preprocessing of laboratory results involved critical tasks such as cleaning data, eliminating duplicates, ensuring consistent units, and handling missing values. These steps are essential to en- sure the reliability of the dataset and to enable ML models to operate effectively. Feature engineering also played a key role in transforming raw data into infor- mative features that the models could use for prediction.

**Feature Engineering** The preprocessing of laboratory results in our research required crucial tasks such data cleaning, duplication removal, unit consistency checking, and handling missing values. These procedures are necessary to guar- antee the dataset’s dependability and to allow ML models to function properly. The conversion of uninformative raw data into informative features that the models could employ for prediction was facilitated in large part by feature engi- neering.

**Model Selection:** Deciding which machine learning (ML) models to use for medical diagnostics is crucial. In our project, we investigated a number of tech- niques, such as XGBoost, Light GBM, Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN). Each of these models has unique strengths and weaknesses, and their performance needed to be evaluated rigorously to deter- mine which one could provide the most accurate diagnoses

**Multi-label Classification:** Medical diagnoses are often not confined to single categories; patients can exhibit multiple co-existing conditions or risk factors. As a result, we opted to explore multilabel classification techniques, allowing our models to predict multiple diagnostic possibilities simultaneously. This approach better reflects the complexity of real-world medical scenarios.

**Evaluation Metrics:** In assessing the models, we employed F1 score and preci- sion as key evaluation metrics. These metrics are particularly relevant in health- care applications where both false positives and false negatives can have serious consequences. High precision and F1 score reflect the ability to minimize both types of errors.

# Motivation

The motivation behind embarking on this machine learning project centered on addressing critical challenges in the field of medical science and healthcare. Several key factors drove our commitment to utilizing ML for diagnosing medical conditions using laboratory results:

## Improving Diagnostic Accuracy

To improve the precision of medical diagnostics was one of the main drivers. Although useful, conventional diagnostic techniques are frequently constrained by the complexity and diversity of diseases as well as the enormous amount of patient data. This data has the potential to be thoroughly analyzed by ML, revealing nuanced connections and patterns that would escape the attention of human therapists. We sought to eliminate misdiagnoses, deliver prompt care, and ultimately enhance patient outcomes by increasing diagnostic accuracy.

## Early Detection and Interventions

To improve the precision of medical diagnostics was one of the main drivers. Although useful, conventional diagnostic techniques are frequently constrained by the complexity and diversity of diseases as well as the enormous amount of patient data. This data has the potential to be thoroughly analyzed by ML, revealing nuanced connections and patterns that would escape the attention of human therapists. We sought to eliminate misdiagnoses, deliver prompt care, and ultimately enhance patient outcomes by increasing diagnostic accuracy.

## Personalized Medicine

The idea of customized medicine, which involves creating treatment programs specifically for each patient based on their distinctive traits, is gaining popularity. Machine learning algorithms can examine enormous datasets, including genetic data, and identify therapies that are more likely to be successful and to cause fewer negative effects. This individualized approach may result in better patient outcomes and experiences.

## Efficiency and Resource Optimization

By maximizing resource allocation and decision-making, machine learning helps streamline healthcare operations. Healthcare facilities may run more produc- tively, shorten wait times, and deploy resources where they are most needed, ultimately raising the standard of care, by automating operations like patient triage, appointment scheduling, and resource allocation.

## Handling Complex and Multi-modal Data

Inherently complicated, medical data frequently consists of a variety of data sources, such as written records, photographs, and numerical data, such as test findings. An extensive study of a patient’s health status is possible thanks to ML’s prowess in processing such multimodal data. This all-encompassing strat- egy is essential for handling complicated medical issues.

## Research and Knowledge Generation

ML can help medical researchers make fresh discoveries and insights. Researchers can find novel risk factors, biomarkers, and treatment possibilities by studying large-scale medical databases, advancing medical knowledge.

# Challenges

Challenges Faced During the Project, Particularly Related to Data Cleaning and Pre-processing:

## Data Quality and Consistency

Raw medical data frequently has problems like missing numbers, outliers, and erratic formatting. To assure the data’s consistency and integrity, significant cleaning and pre-processing were needed. To enable meaningful analysis, miss- ing values had to be imputed, outliers had to be recognized and dealt with appropriately, and the units had to be standardized.

## Data Volume and Scale

Massive healthcare datasets with millions of records can exist. It took effective data processing techniques and computational resources to handle such massive data volumes. A crucial difficulty was ensuring the scalability of the modeling and data pretreatment operations.

## Data Imbalance

Class imbalances are frequent in healthcare datasets, when some illness disorders are more uncommon than others. This disparity may cause model predictions to be skewed. Class imbalances needed to be addressed carefully, which included employing specialist approaches like Synthetic Minority Over-sampling Tech- nique (SMOTE), under-sampling, or over-sampling.

## Feature Engineering

Building powerful machine learning models requires taking the raw data and creating informative features. It was difficult to choose which features to include, how to encode categorical variables, and how to transform continuous variables. To choose pertinent features, domain knowledge was necessary.

## Temporal Data Handling

In medical analysis, time-series data, such as sequential test results, frequently play an important role. Specialized preparation methods were necessary for han- dling temporal data, aligning timestamps, and selecting acceptable time intervals for analysis.

## Regulatory Compliance

Compliance with ethical standards and healthcare legislation was a continual worry. To avoid moral and legal ambiguities, it was crucial to make sure that model development and data preprocessing complied with legal and regulatory requirements.

## Domain Expertise

Due to the interdisciplinary nature of healthcare and machine learning, data scientists and domain specialists in the field of medicine must work closely to- gether. It was a constant task to bridge the gap between these disciplines and make sure that pre-processing was in line with medical understanding.

# Literature Review

## Application of ML in Medical Diagnosis

Due to its capacity to interpret complicated and multidimensional data, ma- chine learning has been increasingly used in medical diagnostics. Researchers have looked into numerous ML algorithms in the context of laboratory findings for objectives including disease prediction, risk assessment, and early detection. Machine learning models are being used fro early detection of prevalent diseases

namely hypertension and diabetes. ML is being used to diagnose diseases with considerable accuracy using vital signs or lab results alone. In certain cases, ML models have been able to achieve a higher accuracy than a trained physician using the same information. [3] uses tree ensembles and DNN for disease predic- tion using lab results. They use 86 lab results to predict 39 labels. A total of 88 features are used including age and sex. XGBoost and LightGBM tree ensembles are used for predictions. XGBoost shows comparatively better results. The pa- per uses multi class classification technique. A tee ensemble using XGBoost and LightGBM is used to further improve the results. For the five most prevalent diseases, the optimized ensemble model had a prediction accuracy of 92 percent and an F1-score of 81 percent. The prediction abilities and patterns for disease classification between the deep learning and ML models were different.

* + 1. focuses on predicting diabetes and cardiovascular diseases. This paper also uses weighted ensembles to for predictions. Cross-fold validation is used for selection of hyper-parameters. [2] explains the approach of hybrid feature engineering of medical data via variational autoencoders.

The academic literature at large suggests that weighted ensembles provide better results in classification problems specially when dealing with large number of features. K-fold cross validation is a convenient way to select the optimised hyper-parameters.

The best results observed in [3] were achieved via DNN. Conceptually, it also makes sense to apply deep learning techniques to predict diseases out of a number of labels specially when dealing with large data sets.

## Challenges in Data Preprocessing

Data preparation is a common issue when using ML to make medical diagnoses based on laboratory results. As a result of problems including missing values, data imbalance, and privacy concerns, raw medical data frequently needs con- siderable cleaning, standardization, and feature engineering.

**Data Imbalance** It might be difficult to manage datasets that are unbalanced. To overcome class imbalances in medical data, researchers have used strategies like oversampling, undersampling, and cost-sensitive learning.

**Privacy and Security** The importance of adhering to data privacy laws like HIPAA cannot be overstated. Techniques for protecting sensitive patient data during data preparation, such as anonymization and encryption, have been de- veloped.

## Model Selection and Evaluation

Accurate diagnosis depends on the selection of suitable ML algorithms. In the context of medical diagnosis, comparative studies have assessed the performance of numerous models, including logistic regression, support vector machines, and

ensemble approaches. To evaluate model performance, evaluation metrics includ- ing F1 score, precision, and recall are frequently utilized.

## Interpretability and Clinical Adoption

For ML models in the medical field to earn the trust of healthcare practitioners, interpretability is crucial. Techniques for explaining model predictions have been studied in order to make them easier to interpret and use in clinical settings.

# Methodology

## Data Preprocessing

**Data Cleaning** Deal with problems in the laboratory results data such missing numbers, outliers, and inconsistent formatting. Make use of the proper tools to impute missing values.Imputing null values with median values gave the best results.

**Data Standardization** We ensured consistency in units and formats across laboratory results to facilitate meaningful analysis.

**Data Integration** We extracted data from two different SQL tables. We had to merge them in to a single table with a unique identifier. A patient ID was unique in a particular practice. Both entities were combined to create a unique identifier which was used to merge data from different sources.

**Data Labeling** We labelled the data for supervised learning, ensuring accurate annotations for each medical condition or outcome of interest.

## Feature Engineering

From the preprocessed data, create features that are pertinent. Aggregating time-series data, encoding categorical variables, and converting continuous vari- ables is necessary to do this. When choosing informative features that are likely to be connected to medical diagnosis, we took domain expertise into considera- tion.

## Data Splitting

Create training, validation, and test sets from the dataset. As a result, distinct data subsets can be used for model training, hyperparameter adjustment, and performance assessment.

## Model Selection

We tried out various machine learning models that are appropriate for prob- lems involving medical diagnosis. Logistic regression, support vector machines, decision trees, random forests and gradient boosting methods (XGBoost, Light- GBM) are examples of models used.

## Model Training

We utilized the training dataset and the optimum methods and hyperparameters to train a selection of machine learning models. We implemented methods such as oversampling, undersampling for handling class imbalances present in the labeled dataset.

## Model Evaluation

We assessed the performance of trained models on the validation dataset using relevant evaluation metrics, such as F1 score, precision, recall and hamming loss.

# Results

We achieved best results using XGBoost classifier. The model achieved precision of 61 percent and an F1 score of 36.9 percent. Along with them, a hamming loss of 5 percent was observed. We used 200 features and predicted 39 labels. Reducing the number of labels further reduced the scores. Likewise, training on a larger data set with around more than 80 thousand patients also diminished the performance of the model as the features’ columns became more and more sparse. We also tried KNN classifier, random forest and LightGBM. None of these models achieved an F1 score greater than 35 percent. We also experimented with multi class classification. This approach is theoretically inconsistent as there is high correlation between certain diseases i.e. occurrence of one disease may significantly increase the chances of another disease and this conjecture was also verified by the clinical data as many patients were suffering from a number of various diseases. Multi class classification also showed worse results than multi label classification. Another advantage of using XGBoost is that it prevents overfitting. It is resistant to overfitting despite increasing number of trees and the maximum depth of trees. Greater F1 score was achieved in prediction of labels with more occurrence. Individual F1 score of up to 71 percent was achieved. This shows that increasing larger data set and less sparse features can improve the evaluation metrics.

# Conclusion

In conclusion, the current results show there is potential for application of ML in disease diagnosis but incumbent model needs improvement. The results can

be improved by using knowledge driven techniques for feature engineering. Cur- rently, feature selection was done based on sparsity of features. Selecting features based on domain knowledge can significantly improve the results. However, the precision of diagnosis is 61 percent which is quite encouraging.

# References

1. A. Dinh. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making*, 2019.
2. M. Mahdav. Hybrid feature engineering of medical data via variational autoencoders with triplet loss: a covid-19 prognosis study. *Scientifc Reports*, 2023.
3. D. J. Park. Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Scientifc Reports*, 2021.