Towards Predictive Medical Diagnostics

Improving Diagnostic Accuracy with Aggregated Laboratory Test Data

Bachelor Thesis
Faculty of Science, University of Bern

submitted by

Marc Fuhrer

from Bern, Switzerland

Supervision:

PD Dr. Kaspar Riesen
Institute of Computer Science (INF)
University of Bern, Switzerland

Abstract

Machine Learning has numerous medical applications and provides opportunities to improve diagnostic workflows. This thesis explores the predictive capabilities of laboratory test data from the Swiss BioRef dataset using the XGBoost algorithm. It addresses the challenges posed by the anonymized dataset by introducing a demographic-based aggregation approach. Specifically, this study examines the potential of laboratory test results aggregation based on age and gender to simulate more comprehensive patient data. The primary goal is to assess the predictive performance of individual laboratory tests for 41 common diagnoses in the dataset and to evaluate how the aggregation impacts the resulting accuracy of the model. Binary classification is employed to determine whether measurements indicate a diagnosis or not. The key findings in this work are that the model performance significantly increases when providing the aggregated data. The evaluation reveals that aggregating laboratory test results improves model accuracy by over 10 percentage points for multiple diagnoses. Furthermore, this work provides a framework to help understand the diagnostic value of laboratory test data and intends to lay the groundwork for validating it with real-world datasets.

Acknowledgements

I would like to express my sincere gratitude to PD Dr. Kaspar Riesen for the opportunity to conduct my thesis within his research group and for his supervision. I am also thankful to Dr. Aylin Taştan for introducing me to the dataset and for her dedicated guidance and support during the early stages of my thesis. Furthermore, I want to thank PD Dr. Alexander Benedikt Leichtle and Dr. Tobias Ueli Blatter for granting access to the data and for their prompt and helpful responses to my questions and issues.

Contents

Chapter 1

Introduction

1.1 Background and Motivation

The field of Artificial Intelligence (AI) has seen rapid advancements over the last few years. It has fundamentally transformed numerous domains, including finance, entertainment and medicine [1, 2, 3]. Machine Learning (ML) is a subfield of AI that focuses on creating algorithms that can analyze and learn from data to make predictions or decisions. Supervised Learning is a sub-area of ML and is fundamental for prediction tasks. Especially in healthcare, accurate classification and regression models can improve patient outcomes [4].

One of the most impactful applications of supervised ML models is disease diagnosis and prognosis [5]. The timely and accurate diagnosis of diseases is crucial for improving patient outcomes. An early diagnosis can enhance treatment success for diseases such as cancer, cardiovascular diseases or diabetes. For instance, early detection of breast cancer through mammography significantly reduces mortality rates [6], while identifying cardiovascular risk factors such as high cholesterol prevents life-threatening events [7]. Similarly, early detection in diabetes minimizes complications like neuropathy and kidney damage [8]. Besides that, prompt intervention can help reduce healthcare costs significantly [9].

In the last few years, ML has seen remarkable developments in the medical field [10]. In the early 2000s, algorithms were mainly used to analyze medical images such as X-rays and MRI scans [11]. This is changing because of the growing availability of large medical datasets and the advancements in computing power, which expands the application of ML [12]. Today, researchers are advancing precision medicine by applying ML to incorporate multi-modal data, including epigenetic, metabolic, and radiological information, thereby aiming to enable predictive modeling for patient-

specific diagnostics and therapies [13]. Especially, laboratory tests such as blood counts and urinalysis are an important component in medical decision-making, offering quantitative insights into the health of patients. Because of the numerical and structured nature of laboratory testing, it is well suited for computational methods such as ML.[14]

The increasing availability of medical data presents significant opportunities to improve diagnostic precision and efficiency. However, it also brings notable challenges, including ensuring anonymization, addressing class imbalance, and managing the complexity of integrating diverse data types [15]. Overcoming these challenges is essential to fully harness the potential of ML in advancing medical diagnostics and patient care.

1.2 Research Goals and Objectives

The dataset used in this thesis contains anonymized laboratory blood measurements and is a part of the dataset used for the Swiss BioRef research project [16]. This thesis aims to establish a framework that evaluates the predictive performance of ML models on the most common diagnoses in the Swiss BioRef dataset. The dataset includes multiple measurements for the same patients. Due to the anonymization, the patient measurements cannot be grouped, limiting each entry to a single laboratory test value, age, and administrative gender. This limitation creates challenges in developing precise predictive models because of the lack of sufficient data.

Despite this, this thesis investigates the potential benefits of more comprehensive patient data by artificially aggregating laboratory values from different tests. The goal is to provide a guideline demonstrating how iteratively, including additional laboratory values, influences the accuracy of diagnosis prediction. This analysis evaluates how aggregating patient information impacts the performance of predictive models. Further, laboratory tests that show the best diagnosis prediction accuracy are identified. The resulting guideline illustrates the potential applications of diagnosis prediction for the Swiss BioRef and similar datasets.

1.3 Structure of the Thesis

The structure of this thesis is divided into five main chapters, each building on the previous one. Chapter 2 introduces the foundational concepts of ML, its applications in medicine, and the theoretical principles behind the XGBoost algorithm.

Additionally, related work in the field is reviewed to contextualize this research. Chapter 3 explores the Swiss BioRef dataset, detailing its structure and the methodologies employed for predictive model evaluation and laboratory data aggregation. Chapter 4 presents the experimental findings, highlighting the impact of data aggregation and hyperparameter optimization on predictive accuracy across various diagnoses, and also addresses the limitations of this thesis. Lastly, Chapter 5 summarizes the key insights from this thesis and proposes directions for future research.

Chapter 2

Theory and Background

This chapter provides the theoretical foundations for the later chapters. It revises the most relevant literature and sets the thesis into context. It lays special focus on XGBoost because it is the main ML algorithm used in this work, and discusses its functionality and characteristics in detail.

2.1 Machine Learning in Medicine

This section provides a brief introduction to ML, explaining the main ideas and concepts. Next, the application of ML in medicine is discussed, with a focus on tasks relevant to supervised learning and classification. Finally, the challenges of applying ML to medical data are highlighted.

2.1.1 General Introduction to Machine Learning

ML is a subfield of AI. Its general goal is to find and extract structure within data and fit models that can make predictions, classify information, or uncover patterns. The models learn from the provided data by leveraging statistical methods and computational power, improving their performance on specific tasks by using experiences to make informed decisions on new, unseen data. ML can be separated into three primary categories: Supervised Learning, Unsupervised Learning and Reinforcement Learning.

- *Supervised Learning:* In Supervised Learning, the model is trained with a dataset that contains both inputs (features) and corresponding outputs (labels). The aim is to learn a function that maps features to the correct labels. Examples include classification problems like disease diagnosis or regression problems such as predicting blood sugar levels.

- *Unsupervised Learning:* In Unsupervised Learning, the model works with unlabeled data and seeks to recognize hidden patterns or structures. Typical applications include clustering, where similar data points are grouped together, and dimensionality reduction, where high-dimensional data is projected into a lower-dimensional space for easier analysis.

- *Reinforcement Learning:* Reinforcement Learning is based on a reward system in which an agent learns through interactions with its environment to maximize cumulative rewards. This method is often applied in robotics and optimization of complex decision-making processes.

This thesis employs classification for diagnosis prediction, situating itself within the supervised learning domain.

Classification

Classification is a subarea of Supervised Learning that categorizes data points into predefined classes. For this, a classification model is trained to learn a function that assigns inputs to the correct class label. Classification problems are common in different areas, such as image recognition, spam filtering, and diagnosis prediction. Some of the most widely used classification algorithms are decision trees, support vector machines (SVMs), neural networks, and gradient boosting methods like XGBoost.

The performance of classification models can be evaluated with the help of different metrics:

- *Confusion matrix:* The confusion matrix (Figure 2.1) provides a detailed breakdown of model predictions, showing the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

- *Accuracy:* Accuracy, as shown in Equation 2.1, is the proportion of correctly classified data points (both true positives and true negatives) to the total number of data points.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{FP} + \text{FN}} \tag{2.1}$$

- *Precision, Recall and F1-score:* As shown in Equation 2.2, precision is calculated as the ratio of true positives to the sum of true positives and false positives. Similarly, recall (Equation 2.3) evaluates the model's ability to identify all relevant instances, and the F1-score (Equation 2.4) combines these two

metrics into a harmonic mean.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.3}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.4}$$

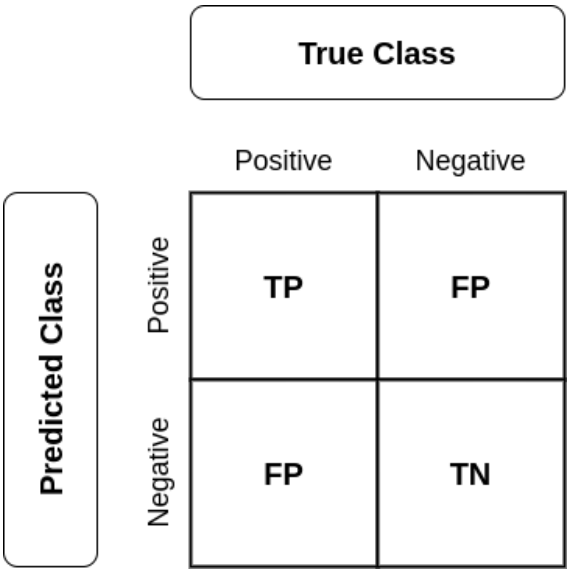

Figure 2.1: Confusion matrix illustrating the distribution of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for the evaluated model.

2.1.2 Applications of Machine Learning in Medicine

There are numerous medical application fields in which ML can be of significance. Obermeyer et al. predict that ML models will take away most of the work of radiologists and anatomical pathologists because their main task is to work with digitized images that can also be given to algorithms [17]. The creation of large image datasets combined with advancements in computer vision will soon enable machines to perform better than humans. Further, they describe that diagnostic precision will be improved as ML algorithms generate differential diagnoses, recommend high-value tests, and reduce unnecessary testing, ultimately enhancing the

efficiency and accuracy of the diagnostic process.

Building on this, Rajkomar et al. explore how ML can extend beyond imaging tasks to augment the broader work of clinicians [18]. They conclude that, by leveraging the vast amounts of data from electronic health records (EHRs) and medical image archives, ML can enhance diagnostic accuracy, aid in early risk detection, and support the development of personalized treatment plans. Additionally, intelligent systems can optimize workflows, enabling clinicians to dedicate more time to direct patient care. Rather than replacing clinicians, ML offers the opportunity to optimize processes and enhance decision-making through data-driven insights.

Challenges of Applying Machine Learning to Medical Data

A prerequisite to successfully applying ML models is the availability of large and high-quality datasets. However, medical data is often complex and presents significant challenges [15]. When working with EHRs, extracting relevant data is difficult because the data can be stored in structured or unstructured formats [19]. Laboratory test results or vital signs are stored in a structured manner within the EHRs and can therefore be easily extracted. Other types of data, such as progress notes, discharge summaries or radiology test results, are stored unstructured in text or images, making them harder to process and analyze without advanced techniques like Natural Language Processing or Computer Vision [20]. While structured data is more straightforward to extract other issues arise like the heterogeneity in data recording practices across healthcare facilities [21]. This complicates the development of robust models that generalize well and show good performance across different datasets.

Another challenge that often arises with medical data is anonymization. It is necessary to protect patient privacy [22]. Due to the removal of unique identifiers, anonymization can make it impossible to link multiple measurements to the same patient, resulting in the loss of temporal or contextual information that is important for the successful application of ML. This lack of connectivity hinders the ability to identify patient-specific trends or relationships. Additionally, anonymization frequently demands the removal of demographic or clinical details that can be helpful for the validation of results and thereby hinder the interpretability of the results. Another common issue in medical datasets is class imbalance, meaning one or more classes have significantly fewer samples than others [23]. An imbalanced dataset can lead to biased ML models that favor the majority class and perform poorly when predicting underrepresented classes like rare diagnoses. Techniques such as

undersampling or SMOTE (Synthetic Minority Oversampling Technique) address this issue and help models better detect rare diagnoses [24].

Related Work

The following examples highlight related work that leverages laboratory test results for the classification of diagnoses or diseases, showcasing the application of ML techniques in medical diagnostics.

Gunčar et al. explore the potential of ML to enhance diagnostic processes for hematologic diseases using laboratory blood test results [25]. For their study, they developed two models using a random forest algorithm. The models were trained with different amounts of parameters, the first with all available 181 parameters and the second with a reduced set of 61 commonly measured parameters. The models were evaluated on two tasks, predicting the most likely hematologic disease and identifying the five most likely diseases. Both models showed promising accuracies of 59% and 57% for the most likely disease and 88% and 86% for the five most likely diseases. The model accuracies were compared to the diagnostic accuracy of hematology specialists and general internal medicine physicians. In the clinical setting, the models performed comparably to specialists and significantly outperformed the general internal medicine physicians. The researchers conclude that a reduced parameter set can effectively capture disease "fingerprints" and ML models have the potential to support physicians in making early and accurate diagnoses.

Park et al. explore the use of ML and Deep Learning to predict diseases using laboratory test results [26]. The data used includes 5,145 cases, encompassing 326,686 laboratory test results and covering 39 classified diseases. For training and evaluation a total of 88 features are used including 86 laboratory tests, age and sex. They individually evaluated XGBoost, LightGBM, and Deep Neural Network (DNN) and combined them as an ensemble model. XGBoost performed the best in predicting the top 5 most likely diseases, achieving an accuracy of 93% and an F1-score of 78%. The LightGBM model performed similarly, reaching an accuracy of 91% and an F1-score of 76%. The DNN model, which was optimized with two hidden layers, achieved an accuracy of 91% and an F1-score of 80% but excelled in classifying specific diseases such as sepsis, scrub typhus, and viral hepatitis. The ensemble model created from the three models showed the best performance with the highest F1-score of 81%. The models were also compared with five physicians, consistently showing superior performance in identifying both the most likely disease and the top five diseases.

The thesis by Gribi utilizes the same dataset as this work [27]. The challenge of anonymization in the Swiss BioRef dataset is addressed by grouping subsequent measurements with matching age and gender, creating an aggregated dataset for analysis. His work focuses on ten specific diagnoses: the four most common in the dataset, three chosen because of their classification potential (Type 2 Diabetes Mellitus, Duration of Pregnancy, and Iron Deficiency Anemia), and three least common diagnoses with more than 1,500 cases. Gribi utilizes unbalanced datasets for his analysis, which reflects the real-world distribution of diagnoses. For each of the ten diagnoses a dataset is created for binary classification and evaluated using XGBoost. The chosen diagnosis is therefore defined as the positive class and all others are treated as the negative class. His results include a maximum F1-score of 78% for the pregnancy diagnosis, with an average F1-score of 39% across all ten diagnoses. The conclusion drawn is that the performances are poor compared to similar tasks in the literature, mainly due to class imbalance, but highlights the dataset's potential for ML applications.

2.2 XGBoost

This section describes the XGBoost algorithm, focusing on its application to binary classification tasks using tree-based boosting. Further, the algorithm's functionality, advanced features, and hyperparameter are detailed below.

2.2.1 XGBoost Algorithm Overview

XGBoost (eXtreme Gradient Boosting) is an ML algorithm designed for efficiency and high predictive performance, first introduced by Chen and Guestrin [28][1]. XGBoost stands out from other gradient boosting algorithms because of its scalability and computational efficiency, making it suitable for large datasets. Given its high accuracy and speed, XGBoost has become a popular choice in various domains, including medicine. For example, there are a number of papers using XGBoost for disease prediction [26, 29, 30, 31].

Gradient Boosting

Gradient boosting is the core concept underlying XGBoost and was originally developed by Friedman [32]. It builds an additive model by sequentially combining weak learners, typically decision trees, to minimize prediction errors iteratively. Weak

[1]The official XGBoost implementation is available at `https://github.com/dmlc/xgboost`.

learners refer to shallow trees that perform slightly better than random guessing. They tend to have high bias and are not very powerful individually. Therefore, each tree is constructed to correct the errors made by the previous ones, progressively improving the model's accuracy.

Given a dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^m$ represents the input features and y_i the target values, gradient boosting constructs an additive model:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i), \tag{2.5}$$

where:

- $\hat{y}_i^{(t)}$ is the prediction at iteration t,
- $\hat{y}_i^{(t-1)}$ is the prediction from the previous iteration,
- $f_t(x)$ is the new decision tree added to reduce the prediction error.

The objective function, which is minimized at each iteration, is defined as:

$$L^{(t)} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{(t)}), \tag{2.6}$$

where ℓ is the loss function. Log-loss is commonly used for binary classification tasks. To optimize $f_t(x)$, the gradient of the loss function with respect to the predictions $\hat{y}_i$ from the previous iteration is calculated:

$$g_i = \frac{\partial \ell(y_i, \hat{y}_i)}{\partial \hat{y}_i}. \tag{2.7}$$

This gradient indicates how the model should adjust its predictions to minimize error. The gradient boosting process efficiently reduces the overall loss by iteratively updating predictions through successive trees.

2.2.2 Advanced Features of XGBoost

XGBoost incorporates advanced features that enhance its performance and usability in real-world applications. These include robust handling of missing data and effective regularization techniques to prevent overfitting.

Handling Missing Data

XGBoost provides a built-in mechanism to handle missing data during training and inference. Instead of discarding incomplete records or resorting to imputation,

XGBoost automatically determines the most suitable branch for missing features when splitting nodes. For each split in a tree, features with missing entries are assigned a default direction, either left or right. During the training process, the algorithm evaluates both possible directions and selects the one that results in the greatest improvement in the model's performance. Specifically, the decision is based on evaluating the loss function for both directions and choosing the direction that minimizes the error. This strategy ensures that missing values are handled effectively, minimizing their impact on the model.

Regularization

XGBoost employs L_1 (Lasso) and L_2 (Ridge) regularization to prevent overfitting, a common challenge in high-dimensional datasets. Regularization adds a penalty term to the objective function, discouraging overly complex models and improving generalization.

The regularized objective function for XGBoost is defined as:

$$L = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \lambda \sum_{k=1}^{T} w_k^2 + \alpha \sum_{k=1}^{T} |w_k|, \tag{2.8}$$

where:

- $\ell(y_i, \hat{y}_i)$ is the loss function (e.g. squared error or log-loss),
- T is the number of leaves in the tree,
- w_k is the weight of leaf k,
- λ is the L_2 regularization parameter, and
- α is the L_1 regularization parameter.

The L_1 term ($\alpha \sum |w_k|$) reduces the complexity of the model by encouraging some weights to become zero, effectively removing unnecessary branches. The L_2 term ($\lambda \sum w_k^2$) prevents the weights from becoming too large, leading to smoother and more stable predictions. Together these techniques help prevent the model from overfitting.

2.2.3 Hyperparameter Tuning

XGBoost is a highly flexible ML algorithm with numerous hyperparameters influencing its performance and behavior. In this thesis, XGBoost is applied specifically

for binary classification tasks, where the goal is to categorize instances into one of two possible classes. Further, the tree-based booster `gbtree` is used. When using `gbtree` for binary classification tasks, hyperparameters of the algorithm control aspects like tree structure, learning process, and regularization. Proper tuning of these hyperparameters is crucial for achieving optimal predictive accuracy and generalization. Table 2.1 outlines the key hyperparameters for the XGBoost classifier with a tree booster.

Hyperparameter	**Description**
`n_estimators`	Number of decision trees.
`max_depth`	Maximum depth of a tree; controls model complexity.
`learning_rate`	Step size shrinkage; balances weight updates between iterations.
`subsample`	Fraction of samples used for training each tree; prevents overfitting.
`colsample_bytree`	Fraction of features used for each tree.
`min_child_weight`	Minimum sum of weights required for child nodes; prevents overfitting.
`gamma`	Minimum loss reduction required for a split; regularizes the tree.
`alpha`	L1 regularization term; adds sparsity to the model.
`lambda`	L2 regularization term; prevents overfitting.
`scale_pos_weight`	Balances positive and negative classes for imbalanced datasets.
`objective`	Specifies the learning task (e.g., `binary:logistic` for binary classification).
`tree_method`	Algorithm for constructing trees (e.g., `hist` for large datasets).

Table 2.1: Key hyperparameters for the XGBoost classifier using `gbtree`.

This thesis uses Bayesian optimization to tune the hyperparameters [33]. Bayesian optimization builds a probabilistic model of the objective function to find the optimal hyperparameter settings. Bayesian optimization is chosen because it has lower computational costs than grid or random search while ensuring robust optimization.

Chapter 3

Dataset and Methodology

This chapter introduces the dataset used and its original background. It also highlights the creation of subsets to create datasets for specific diagnoses. Further, the aggregation approach is explained and the evaluation of the models is discussed.

3.1 Dataset

This section introduces the dataset used in this thesis, sourced from the Swiss BioRef research project. It outlines the dataset's origin, structure, and the features selected for ML analysis.

3.1.1 Swiss BioRef Dataset

This thesis uses a large part of a medical dataset created for the Swiss BioRef research project [16]. Swiss BioRef is a nationwide initiative that aims to provide a standardized framework for calculating and evaluating patient-group-specific reference intervals in laboratory medicine. Patient-group-specific reference intervals are ranges of laboratory test values that incorporate factors such as age, sex, and specific diagnoses to provide a more accurate and individualized interpretation of test results. The whole dataset comprises harmonized laboratory data from four major Swiss hospitals: Inselspital Bern, University Children's Hospital Zurich, CHUV Lausanne, and Swiss Paraplegic Research. This thesis uses only data from the Inselspital Bern.

3.1.2 Structure of the Dataset

The dataset consists of 5,967,847 anonymized laboratory blood test results collected from 186,265 patients. It includes measurements from 39 different laboratory tests

(see Table B.1 in the Appendix), each recording the test performed, the value measured, and patient information such as age and gender. The laboratory tests are encoded using Logical Observation Identifiers Names and Codes (LOINC) [34]. Each entry includes up to five relevant diagnoses, classified according to the 10th Revision of the International Classification of Diseases, German Modification (ICD-10-GM) [35]. In total, each measurement consists of 21 features, as detailed in Table B.2 in the Appendix.

The following uses only nine features of the dataset (see Table 3.1. Features used as patient or case identifiers and features related to laboratory equipment used for the blood tests are excluded. These features are assumed to be irrelevant for predictive modeling and do not hold additional information relevant to ML algorithms. Therefore, the rest of the thesis omits these features.

Labtest	LabResultValue	Diag01	Diag02	...	Diag05	Age	Gender
-	-	-	-	-	-	-	-

Table 3.1: The nine features of the Swiss BioRef dataset assumed relevant for ML.

3.2 Evaluation of the Best Predictors for Diagnoses

The evaluation of the best predictors for diagnoses involves a systematic approach to identify and analyze the most relevant features in the dataset. This begins with the creation of subsets tailored for binary classification, enabling a focused assessment of diagnostic relevance.

3.2.1 Subset Creation for Diagnosis Evaluation

The first step of the analysis involves identifying the most common diagnoses in the dataset. This is achieved by calculating the frequency of each diagnosis across all measurements and aggregating the occurrences from the columns `Diag01` to `Diag05`. A threshold of 2% relative frequency is applied to select diagnoses for further investigation. This process narrows the analysis from over 1,300 unique entries to just 41 diagnoses (see Figure 3.1).

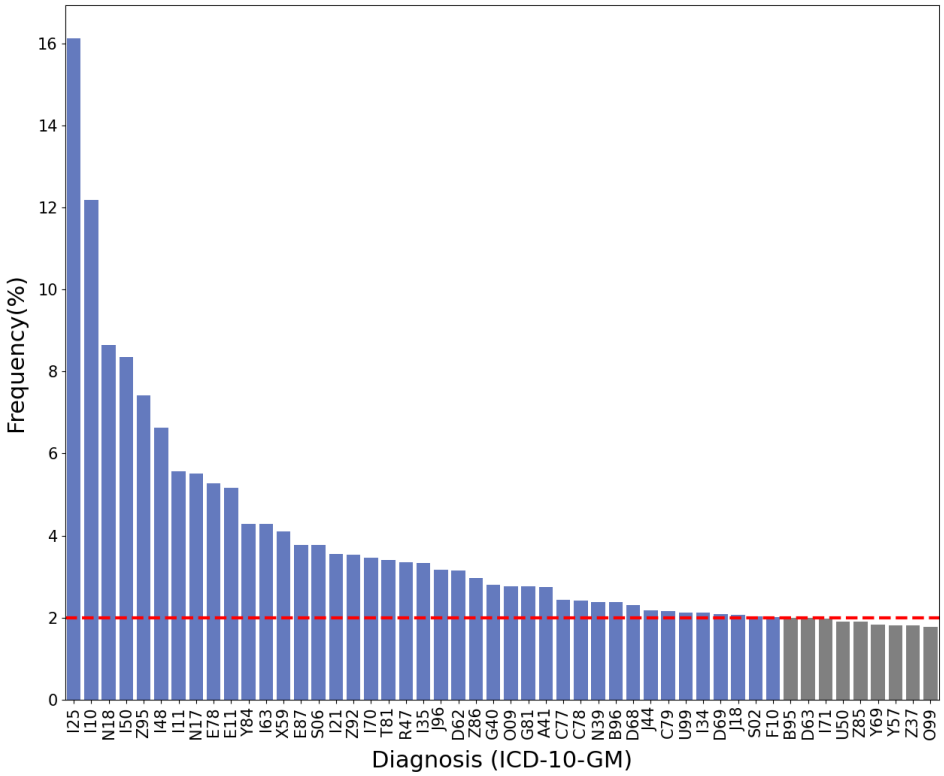

Figure 3.1: Bar plot displaying the 50 most common diagnoses in the Swiss BioRef dataset. Diagnoses with a relative frequency exceeding 2% are represented by blue bars, while those below this threshold are shown in grey. The red dotted line indicates the 2% frequency threshold.

For each of the 41 diagnoses exceeding the threshold, separate subsets are created for each of the 39 distinct laboratory tests to enable binary classification. First, a copy of the original dataset is created and filtered to include only measurements corresponding to the selected laboratory test. Next, a new column named `Diag` is added. If the target diagnosis appears in any row, `Diag` is assigned a value of 1; otherwise, it is assigned 0. The columns `Diag01` through `Diag05` and `Labtest` are then removed. In the final step, undersampling is applied to obtain balanced subsets and reduce the number of undiagnosed cases. This process labels each measurement as diagnosed (1) or undiagnosed (0), yielding balanced subsets for each of the 41 diagnoses in combination with all 39 laboratory tests and resulting in a total of 1599 subsets.

3.2.2 Performance Evaluation Using XGBoost

The XGBoost algorithm is used to evaluate the predictive performance of the individual laboratory tests for the diagnoses. Therefore, the subsets created in the previous step are evaluated using XGBoost with standard parameters. Using standard parameters ensures a consistent baseline and avoids variability from hyperparameter tuning. This approach isolates the inherent predictive value of each laboratory test for each diagnosis.

For the evaluation, each subset is split into a training set (70%) and a test set (30%). Subsets with fewer than 1,000 measurements are excluded to ensure reliable results. Since the subsets are balanced, accuracy is chosen as the primary evaluation metric to assess how well the model distinguishes between diagnosed and undiagnosed cases. The predictive performance of different laboratory tests is compared for the same diagnosis, with higher accuracy indicating stronger predictive relevance.

The results from this step provide the foundation for the data aggregation discussed in the next section. By identifying the best-performing tests, the aggregation can be performed on the base of quantitative insights.

3.3 Data Aggregation

This section explains the methodology for combining laboratory test results to create enriched subsets. The aggregation process uses age and gender to merge laboratory values from multiple measurements while ensuring biological plausibility.

3.3.1 Demographic-Based Data Aggregation

Multiple measurements in the dataset must belong to the same patient since it consists of 5,967,846 laboratory blood test results from 186,265 patients. But because the dataset, as discussed in Section 3.1, is anonymized it is impossible to link multiple laboratory measurements to the same patient. Therefore, a data aggregation approach based on demographics is implemented to address this issue and create subsets with more comprehensive information. The approach utilizes the two available demographic features, age and gender, from the dataset to combine laboratory test results across measurements. This method ensures that the aggregated information remains biologically plausible and preserves demographic consistency. By creating the enriched subsets, it becomes possible to evaluate the

performance of binary classification in a setting that approximates having complete patient information.

3.3.2 Creation of Aggregated Subsets

The first steps to create aggregated subsets is to identify the laboratory test with the highest predictive accuracy for each diagnosis, as determined in Section 3.2. The subset of the best-performing laboratory test is selected as the base subset. The base subset serves as the foundation for the further aggregation. It contains the features `Age`, `Gender`, `LabResultValue` and `Diag`. The feature `LabResultValue` is renamed to `LabResultValue0`.

The aggregation process is performed in a class-specific manner. For diagnosed cases, only diagnosed measurements are considered for aggregation and vice versa. Mathematically, let D_0 denote the base subset containing the feature `LabResultValue0`, and let D_i represent the subsets of laboratory tests ranked by accuracy. For each additional laboratory test D_i, a new feature $\texttt{LabResultValue}_\texttt{i}$ is introduced. Aggregation is performed by matching entries in D_0 and D_i based on the demographic attributes age (A) and gender (G). Specifically, for an entry $x \in D_0$, the matching set M_i is defined as:

$$M_i = \{y \in D_i \mid A(x) = A(y),\, G(x) = G(y)\} \tag{3.1}$$

A fallback mechanism is applied if M_i is empty. In this case, entries from D_i are matched with D_0 such that the gender (G) remains the same, and the age (A) of the matching entries is within a range of ± 2 years of the age of x. The updated matching set is defined as:

$$M_i^{\text{fallback}} = \{y \in D_i \mid |A(x) - A(y)| \leq 2,\, G(x) = G(y)\} \tag{3.2}$$

From the matching set M_i (or M_i^{fallback} if M_i is empty), a laboratory value is sampled randomly and assigned to $\texttt{LabResultValue}_\texttt{i}$ in D_0. This process is repeated iteratively, adding new features $\texttt{LabResultValue}_\texttt{i}$ until n selected laboratory tests are aggregated.

Illustrative Example for Aggregation Process

This following example illustrates the demographic-based aggregation for a Diagnosis X. The base subset D_0 contains one measurement from the best-performing laboratory test (Test A). The objective is to add a measurement from the second-

best performing laboratory test (Test B) using the defined aggregation process. The base subset D_0 contains the following entry:

Age	**Gender**	**LabResultValue0 (Test A)**	**Diag**
50	Male	7.5	1

Table 3.2: Base subset D_0 with one measurement from the best-performing laboratory test (Test A) for Diagnosis X.

The subset D_1 contains measurements from the second-best performing laboratory test for Diagnosis X:

Measurement	**Age**	**Gender**	**LabResultValue (Test B)**	**Diag**
Measurement 1	50	Male	8.2	1
Measurement 2	52	Male	7.8	1
Measurement 3	48	Male	7.9	1
Measurement 4	50	Female	8.1	1
Measurement 5	50	Male	8.0	1

Table 3.3: Subset D_1 contains multiple measurements from the second-best performing test (Test B) for Diagnosis X with potential matching entries.

The matching process proceeds as follows:

1. Exact Match:
    - Filter D_1 for entries with the same age and gender as the measurement in D_0 (Age = 50, Gender = Male).
    - Result: Measurements 1 and 5 match.
2. Fallback Match:
    - If no exact match is found, filter D_1 for entries with the same gender and an age within ± 2 years of the base measurement.
    - Result: Measurements 2 and 3 satisfy the fallback criteria (Age = 52 and Age = 48, Gender = Male).
3. Sampling:
    - Randomly sample one value from the matching set. For this example, Measurement 1 is chosen with LabResultValue = 8.2.

The base subset D_0 is updated to include the new measurement from Test B:

Age	Gender	LabResultValue0	LabResultValue1	Diag
50	Male	7.5	8.2	1

Table 3.4: Updated base subset D_0 after aggregation.

This process is repeated for all measurements in D_0 and for additional laboratory tests to create a dataset with multiple laboratory test values for each diagnosis.

By aggregating laboratory values in this manner, the dataset simulates comprehensive patient information. This stepwise approach enables a systematic evaluation of how the inclusion of additional laboratory tests affects predictive performance, providing a framework for understanding the potential impact of richer datasets in medical diagnostics.

3.4 Evaluation of Aggregated Subsets

To study the effect on the model's performance of subsets with more laboratory test values, we create the subsets step by step. We start with one test and add one test at a time, up to 20 tests for each diagnosis. After adding each test, we evaluate the model with the subset to see how it affects accuracy. In order to reduce randomness and get reliable results, the process is repeated seven times for each diagnosis. We evaluate every subset twice: once with XGBoost using default hyperparameters (see Table B.4) and once with Bayesian optimization to adjust the hyperparameters.

Each time the aggregated subset is evaluated, the data is split into a training set (70%) and a test set (30%). XGBoost is then used to evaluate classification performance through two protocols. First, default parameters are applied to establish a baseline for predictive performance. Second, Bayesian optimization is employed to fine-tune the hyperparameters, using 3-fold cross-validation to evaluate model performance during the tuning process. A detailed overview of the hyperparameters used for tuning, including their descriptions and ranges, is provided in Table 3.5.

The results of this evaluation serve two main purposes. First, they quantify the diagnostic value added by aggregating additional laboratory tests, highlighting the potential for richer datasets to improve predictive performance. Second, they provide insights into the effectiveness of hyperparameter optimization, demonstrating its role in enhancing classification accuracy. This thesis establishes a framework for understanding the potential of performing diagnosis prediction with blood mea-

surements by systematically evaluating aggregated subsets.

Hyperparameter	**Description**	**Range/Value**
`n_estimators`	Number of decision trees	100–300
`max_depth`	Maximum depth of each tree	3–7
`learning_rate`	Step size shrinkage to prevent overfitting	0.01–0.2
`subsample`	Fraction of samples used for tree building	0.6–1.0
`colsample_bytree`	Fraction of features used per tree	0.6–1.0
`alpha`	L1 regularization weight to enforce sparsity	0–5
`lambda`	L2 regularization weight to reduce overfitting	0–10

Table 3.5: Overview of the hyperparameters used for XGBoost with aggregated subsets, including their descriptions and the corresponding ranges.

Chapter 4

Results and Discussion

4.1 Evaluation Results

This section presents the experimental results obtained from the evaluation process. It starts with evaluating the predictive performance of laboratory tests on individual diagnoses. It continues with the impact of demographic-based aggregation and hyperparameter tuning on the model performance. The last part addresses the limitations of the approach.

4.1.1 Predictive Performance of Laboratory Tests

The evaluation of individual laboratory tests is conducted as described in Section 3.2.2, utilizing the XGBoost algorithm with standard parameters to ensure consistency. For each diagnosis, the accuracy of subsets corresponding to individual laboratory tests is measured, with higher accuracy indicating stronger predictive potential. All subsets that contain fewer than a 1,000 measurements are excluded from the analysis.

The evaluation results are visualized in Figure 4.1 as a heatmap. Diagnoses are represented on the x-axis by their ICD-10-GM codes, arranged alphabetically to reflect the hierarchical structure of the ICD-10-GM classification, where codes sharing the same initial letter belong to the same diagnostic category. Laboratory tests are displayed on the y-axis and can be identified by their LOINC codes. The laboratory tests are sorted numerically.

For the diagnosis O09 (Supervision of high-risk pregnancy), several laboratory tests demonstrate strong predictive accuracy. This indicates that specific laboratory val-

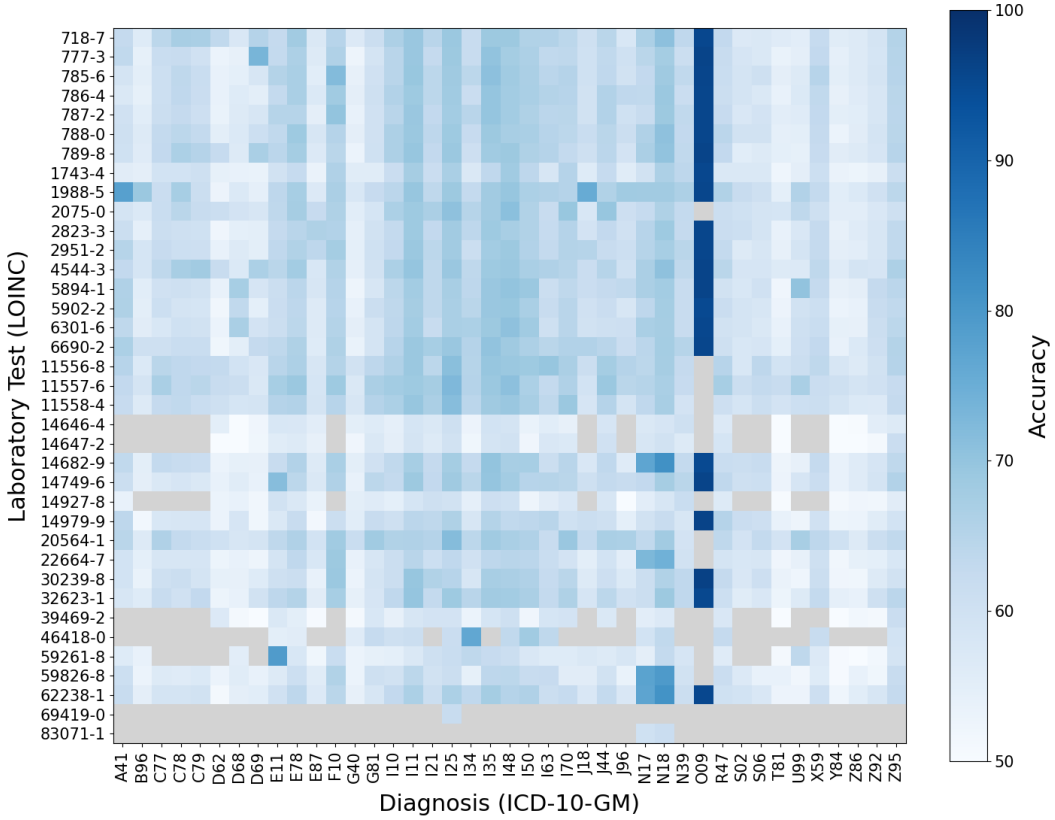

Figure 4.1: Heatmap of accuracies for each laboratory test (y-axis, LOINC codes) across the most common diagnoses (x-axis, ICD-10-GM codes) in the Swiss BioRef dataset. Grey fields indicate subsets excluded due to insufficient data. Laboratory tests 68438-1 (25-Hydroxyvitamin D3+25-Hydroxyvitamin D2) and 20448-7 (Insulin in Plasma or Serum) are omitted as no subset contains more than 1,000 measurements.

ues are reliable indicators for identifying high-risk pregnancies. Additionally, the diagnoses N17 (Acute kidney failure) and N18 (Chronic kidney disease) show similar predictive performance for the same laboratory tests. The laboratory tests that achieve high predictive accuracy for both kidney-related diagnoses include 59826-8 (Creatinine in Blood), 22664-7 (Blood Urea in Serum or Plasma), 62238-1 (Estimated Glomerular Filtration Rate), and 14682-9 (Serum Creatinine in Serum or Plasma). The diagnoses C77 (Secondary and unspecified malignant neoplasm of lymph nodes), C78 (Secondary malignant neoplasm of respiratory and digestive

organs), and C79 (Secondary malignant neoplasm of other sites) share common laboratory tests with strong predictive performance. Specifically, 789-8 (Erythrocytes in Blood), 718-7 (Hemoglobin Concentration in Blood), and 4544-3 (Hematocrit in Blood) are particularly effective in predicting these diagnoses.

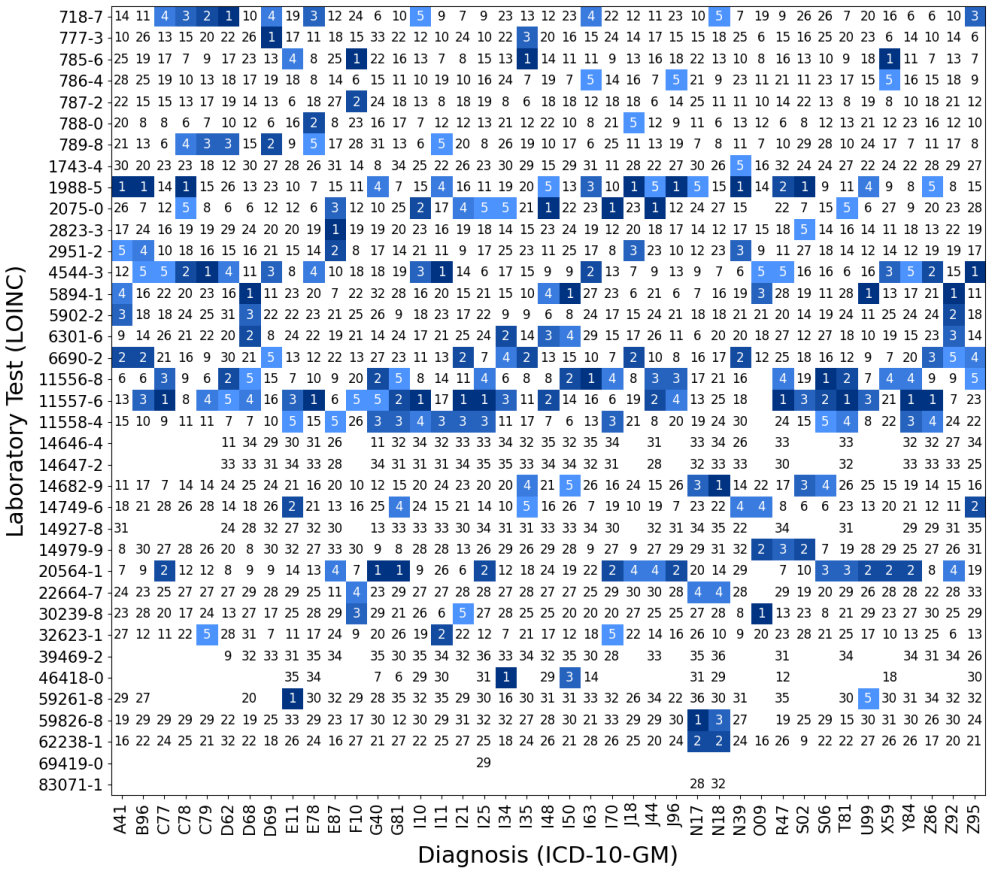

Figure 4.2: Ranking of the top laboratory tests (y-axis, LOINC codes) with the highest accuracies for each diagnosis (x-axis, ICD-10-GM codes). The top five tests are highlighted. Empty spaces indicate subsets excluded due to insufficient data. Laboratory tests 68438-1 (25-Hydroxyvitamin D3+25-Hydroxyvitamin D2) and 20448-7 (Insulin in Plasma or Serum) are omitted as no subset exceeds 1,000 measurements.

Figure 4.2 ranks the laboratory tests for the diagnoses. For each diagnosis, the five laboratory tests with the highest predictive accuracy are highlighted. Notably, certain laboratory tests are good predictors across multiple diagnoses. 11557-6 (Carbon Dioxide Partial Pressure in Blood), 11556-8 (Oxygen Partial Pressure in

Blood), 4544-3 (Hematocrit in Blood), 1988-5 (C-Reactive Protein in Serum or Plasma), and 20564-1 (Oxygen Saturation in Blood) consistently perform well and exhibit an average rank below 10 across all diagnoses. The figure also indicates that the laboratory test 11557-6 (Carbon Dioxide Partial Pressure in Blood) is among the best predictors for the diagnoses I10 (Essential hypertension), I21 (Acute myocardial infarction), I25 (Chronic ischemic heart disease), I34 (Nonrheumatic mitral valve disorders) and I50 (Heart failure). These results show that specific laboratory tests exhibit similar predictive performance for related diagnoses. This could reflect the underlying physiological or pathological commonalities.

4.1.2 Effect of Demographic-Based Aggregation

Figure 4.3 illustrates the impact of demographic-based aggregation on the model accuracy for 15 diagnoses of the 41 most common. The line plot displays the progression of accuracy as additional laboratory test values are aggregated, starting from a single laboratory test and increasing incrementally up to 20 laboratory tests. This approach evaluates how the inclusion of more predictors affects the model's performance. Results for the remaining diagnoses are provided in Appendix A.

The line plot consistently shows an upward trend, indicating that aggregating additional laboratory tests generally enhances accuracy across all diagnoses. Certain diagnoses, such as N18 (Chronic kidney disease), N17 (Acute kidney failure), and I50 (Heart failure), exhibit substantial improvements of over 10 percentage points in accuracy. On the other hand, diagnoses like E11 (Type 2 diabetes mellitus) and I34 (Nonrheumatic mitral valve disorders) show only marginal gains, suggesting that these conditions may be more challenging to predict even with more test values available. It could also suggest that the additional aggregated tests provide limited information and are not particularly relevant to these diagnoses.

Many diagnoses significantly improve when the first few laboratory tests are added. For instance, E87 (Disorders of fluid, electrolyte, and acid-base balance), I50 (Heart failure), X59 (Exposure to unspecified factors causing external injury), and S02 (Fracture of skull and facial bones) exhibit marked increases in accuracy with the addition of the first few laboratory tests. After these initial improvements, the accuracy starts to plateau and shows only minimal further improvements. This behavior can be attributed to the method by adding laboratory tests based on their ranking, as illustrated in Figure 4.2. The laboratory tests with the highest predictive value are incorporated first, contributing significantly to the accuracy

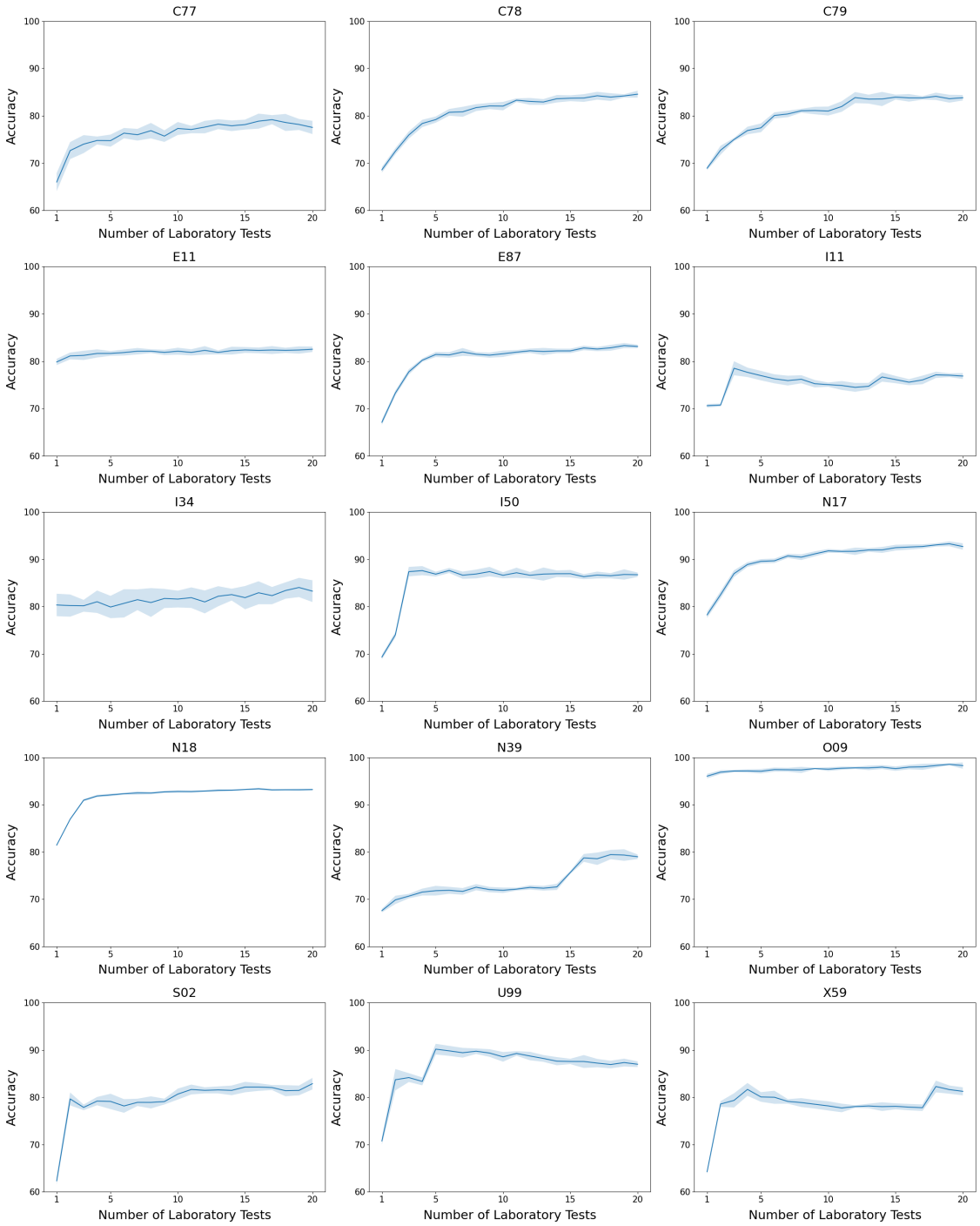

Figure 4.3: Line plots illustrating the progression of model accuracy as additional laboratory tests are aggregated for 15 selected diagnoses (x-axis: number of laboratory tests, y-axis: accuracy). Each line represents the mean accuracy for a diagnosis, with shaded areas indicating the standard deviation.

improvements. Conversely, tests with lower predictive accuracy are added later, which explains the eventual plateau in performance as their contribution becomes negligible. However, some diagnoses deviate from the trend to first improve and

then plateau. For example, X59 (Exposure to unspecified factors causing external injury), U99 (Medical surveillance and observation cases), and N39 (Other disorders of the urinary system) exhibit a sudden increase in accuracy after previously plateauing. This "jump" could suggest that certain tests added later in the sequence may actually be better predictors than they are ranked.

4.1.3 Impact of Hyperparameter Tuning

Figure 4.4 presents the accuracies achieved when aggregating laboratory tests from 1 to 20, comparing results obtained with and without hyperparameter tuning using Bayesian optimization. For the majority of diagnoses, the performance of the model without hyperparameter tuning closely approximates that of the model optimized via Bayesian methods. The figure shows 15 of the 41 analyzed diagnoses, the remaining results are provided in Appendix A.

On average, across all diagnoses and iterations, the XGBoost model tuned with Bayesian optimization demonstrated a marginally higher accuracy, with an improvement of approximately 0.7 percentage points compared to the model without hyperparameter tuning. The most notable performance difference was observed in the diagnosis I34 (Nonrheumatic mitral valve disorders), where the optimized model achieved an average accuracy increase of 1.6 percentage points over 20 iterations. Conversely, the smallest performance difference is found for the diagnosis O09 (Supervision of high-risk pregnancy), where the Bayesian-optimized model outperforms the unoptimized model by only 0.2 percentage points.

4.1.4 Limitations

The most significant limitation of the results lies in the aggregation of laboratory test values. Due to the anonymization of the dataset, the aggregation is performed on the demographic attributes age and gender rather than linking tests to the same patient. While this approach simulates richer subsets and provides valuable insights, it is important to note that the results are only indicative. They must be validated with datasets where all laboratory tests unequivocally belong to the same patient to ensure the practical relevance of the findings.

Another notable limitation is that the diagnoses are assigned prior to the laboratory tests instead of being derived from them. The diagnoses are included in the dataset to indicate the patient's association with a specific group, as the data were originally collected to establish reference values [16]. Consequently, diagnoses

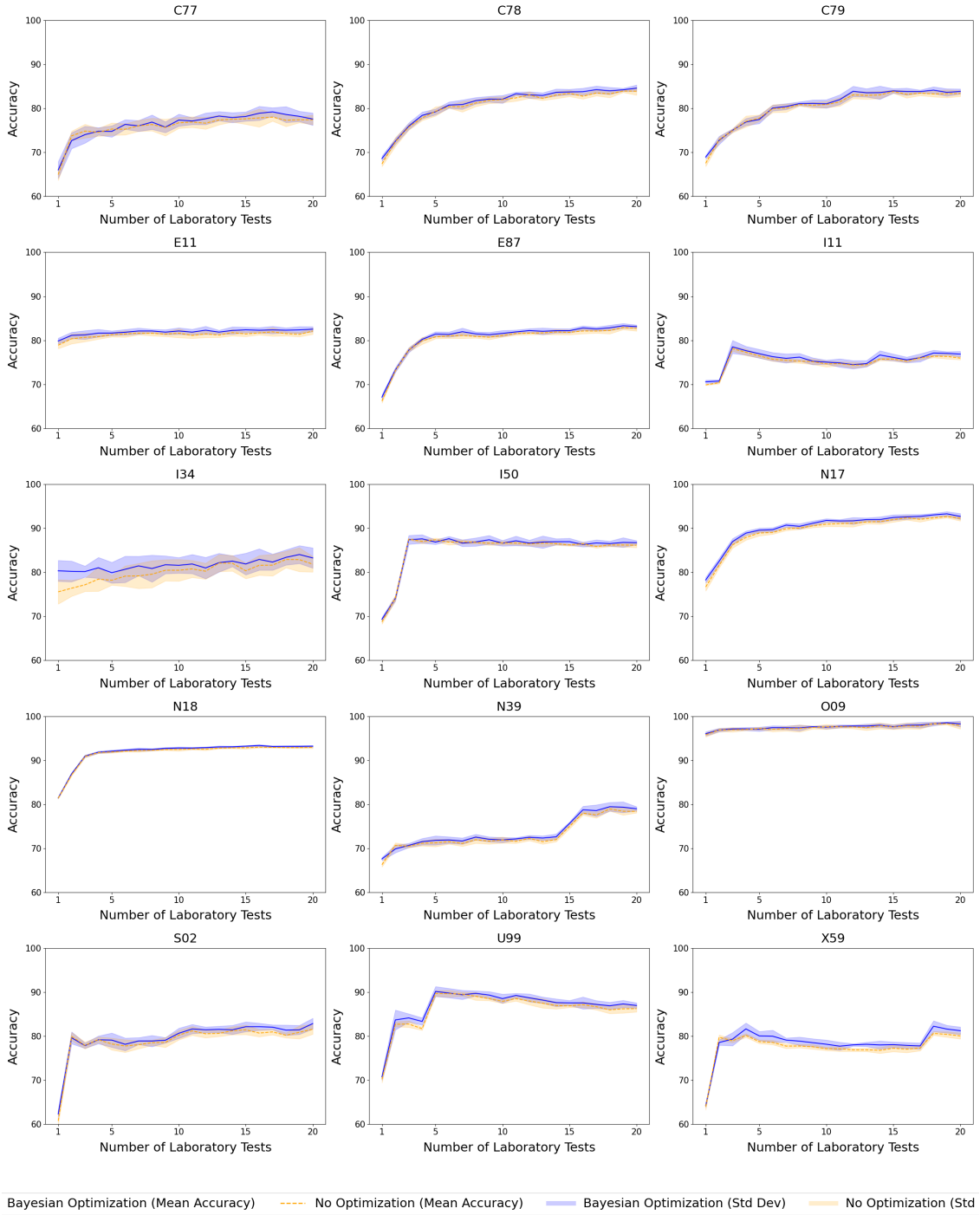

Figure 4.4: Line plots comparing the accuracy achieved when aggregating laboratory tests (x-axis: number of laboratory tests, y-axis: accuracy) for 15 diagnoses, with and without hyperparameter tuning using Bayesian optimization. Each line represents the mean accuracy for a diagnosis, with shaded areas indicating the standard deviation.

serve as metadata about the measurements rather than outcomes of the laboratory tests. This challenges the assumption of a direct statistical relationship between the laboratory results and diagnoses.

Another major limitation concerns the balancing of the subsets derived from the original dataset. The original subsets are like most medical datasets naturally imbalanced, meaning there are a lot more of undiagnosed measurements than diagnosed. For this evaluation, balanced subsets were created to ensure a fair evaluation of model performance. However, such balancing does not reflect real-world distributions. If these methods were applied to the imbalanced subsets, the performance might decline, and accuracy alone would no longer suffice as a reliable evaluation metric. In imbalanced datasets, metrics like precision, recall, or the F1-score are essential to evaluate the model's effectiveness.

Finally, computational constraints posed a limitation. Hyperparameter tuning with Bayesian optimization was intentionally limited to ensure feasibility across 20 iterations for each of the 41 diagnoses. This restricted depth of tuning may have affected the model's optimal performance. A more extensive and thorough tuning process could yield better results.

Chapter 5

Conclusions and Future Work

5.1 Conclusion

The primary objective of this thesis was to evaluate the potential of laboratory blood test values for diagnosing medical conditions using binary classification. The thesis explores how combining laboratory tests can enhance diagnostic accuracy by employing the XGBoost algorithm, a demographic-based data aggregation approach and hyperparameter tuning using Bayesian optimization. Constructing derived subsets from an anonymized dataset provides a structured framework for simulating patient data and enables the assessment of individual and aggregated laboratory test performance. Our key findings are the identification of laboratory tests with high predictive accuracy for specific diagnoses and how the aggregation of different test values improves the model performance.

The evaluation shows that the accuracy achieved by the XGBoost model improves significantly with the aggregation of additional laboratory test values. For instance, the diagnoses N18 (Chronic kidney disease) and N17 (Acute kidney failure) show an accuracy of over 10 percentage points when key laboratory tests are aggregated. In particular, adding the first few laboratory tests, which have the highest predictive potential when evaluated individually for a diagnosis, enhances the performance. Hyperparameter tuning further optimizes the model's accuracy, though the improvement is modest, with an average of 0.7 percentage points across all diagnoses and iterations.

Despite the promising results, they must be interpreted cautiously due to several limitations. Although the aggregation of laboratory tests simulates real-world conditions, it falls short of fully capturing the complexity of actual patient data, as it relies on randomness and assumptions about the relationships between laboratory

tests and diagnoses. As a result, these findings should be viewed as indicative and require validation using non-anonymized datasets with true patient-level records. Another limitation stems from the balancing of subsets for binary classification. Naturally, undiagnosed cases far outnumber diagnosed ones in medical datasets, but balancing these subsets may have artificially inflated the model's classification accuracies. Without balancing, alternative metrics such as precision, recall, or the F1-score would offer a more nuanced and realistic evaluation of the model's performance.

These findings underscore both the potential and the challenges of applying ML to medical diagnostics. While the results emphasize the value of leveraging laboratory test data for predictive modeling, they also emphasize the need for further validation and adaptation to real-world datasets to ensure practical applicability.

5.2 Future Work

The results obtained in this thesis show the potential of ML for predicting medical diagnoses, yet further validation is necessary. Future studies should cross-check these findings using datasets where the laboratory tests definitively belong to the same patient. Working with such datasets would circumvent the reliance on randomness and aggregation, thus providing a clearer understanding of the true predictive power of laboratory tests. Additionally, comparing the rankings of laboratory tests identified in this study with insights from medical practitioners would offer valuable references. Verifying whether the laboratory tests that contribute significantly to the prediction of certain diagnoses align with the tests typically relied upon by medical professionals would enhance the practical relevance of these findings.

Another direction for future work involves exploring other ML models and conducting more extensive hyperparameter tuning to compare their performance with XGBoost. Evaluating different algorithms, such as neural networks or ensemble models, could provide insights for improving diagnostic predictions. Additionally, preprocessing the dataset with outlier detection could help refine the data before applying ML algorithms. Since medical datasets are often affected by extreme values, outlier detection can enhance the robustness and reliability of the models.

Future research can also explore the effect of balanced subsets compared to unbalanced ones. For this, the current approach can be replicated but with unbal-

anced sets. The model's performance could then be assessed using metrics such as F1-score, precision, and recall rather than relying solely on accuracy. This would provide a better understanding of the model's behavior when addressing the additional challenge of unbalanced data. Since data imbalance is common in medical datasets, evaluating the model under these conditions can show insights into the practical applicability.

Finally, this thesis only considers the 41 most common diagnoses in the dataset, but future research can expand the scope to include more diagnoses, especially rare diagnoses. Analyzing and predicting rare diagnoses is often more challenging due to insufficient data and overlapping clinical presentations with more common conditions. This overlap can lead to misdiagnoses, making it harder for ML models to distinguish between rare and common diagnoses.

Appendix A

Additional Figures of Results

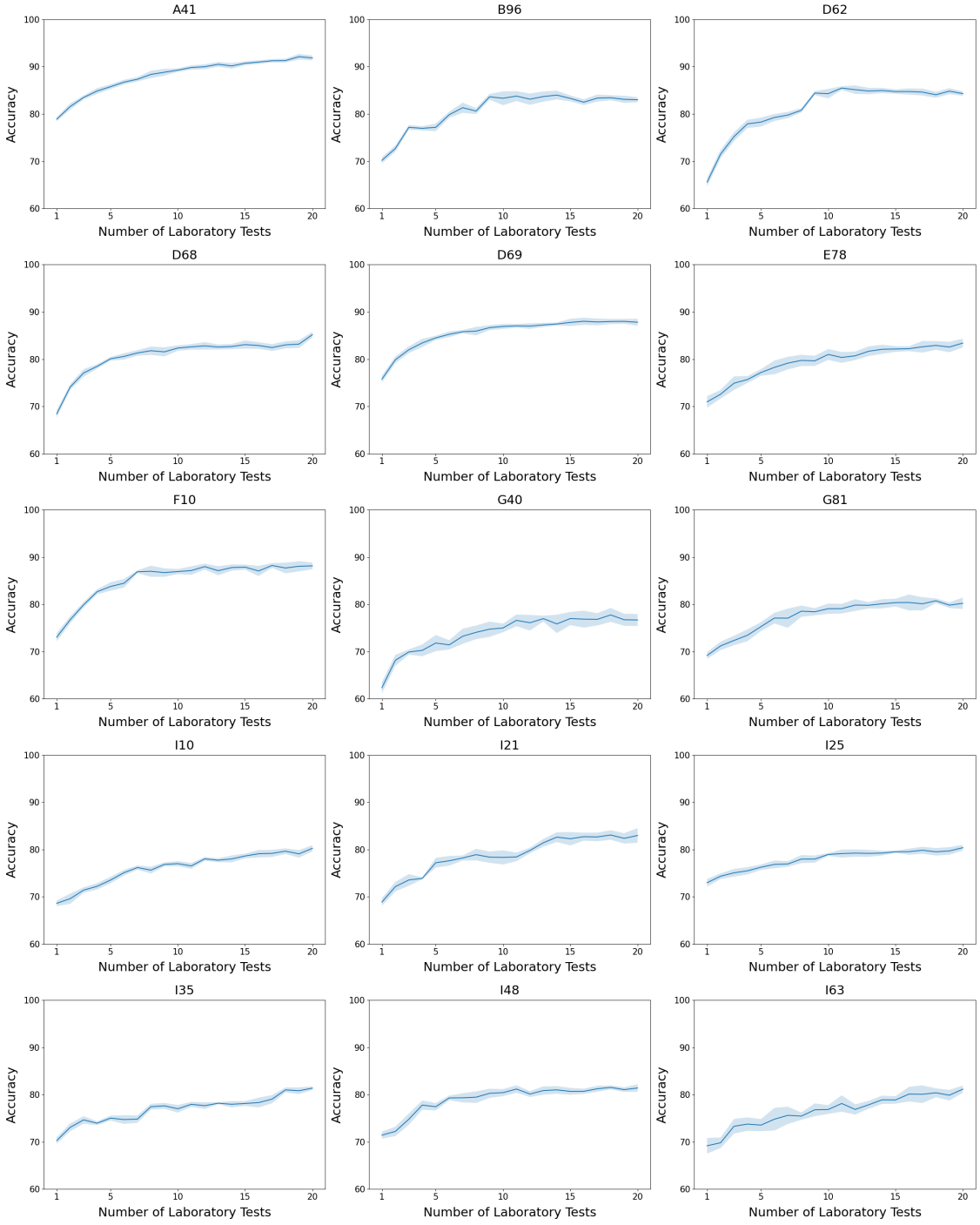

Figure A.1: Line plots illustrating the progression of model accuracy as additional laboratory tests are aggregated for 15 selected diagnoses (x-axis: number of laboratory tests, y-axis: accuracy). Each line represents the mean accuracy for a diagnosis, with shaded areas indicating the standard deviation.

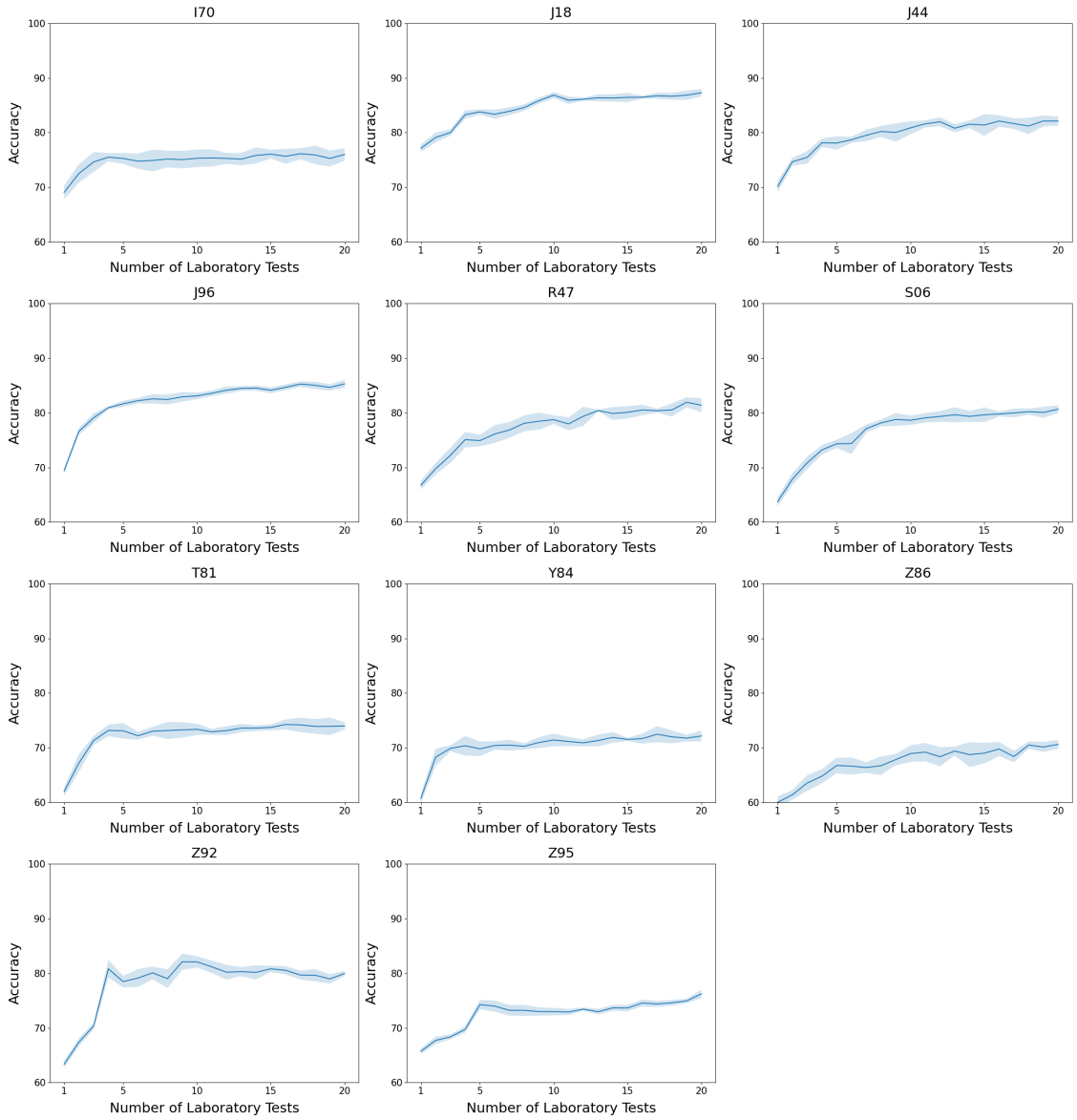

Figure A.2: Line plots illustrating the progression of model accuracy as additional laboratory tests are aggregated for 11 selected diagnoses (x-axis: number of laboratory tests, y-axis: accuracy). Each line represents the mean accuracy for a diagnosis, with shaded areas indicating the standard deviation.

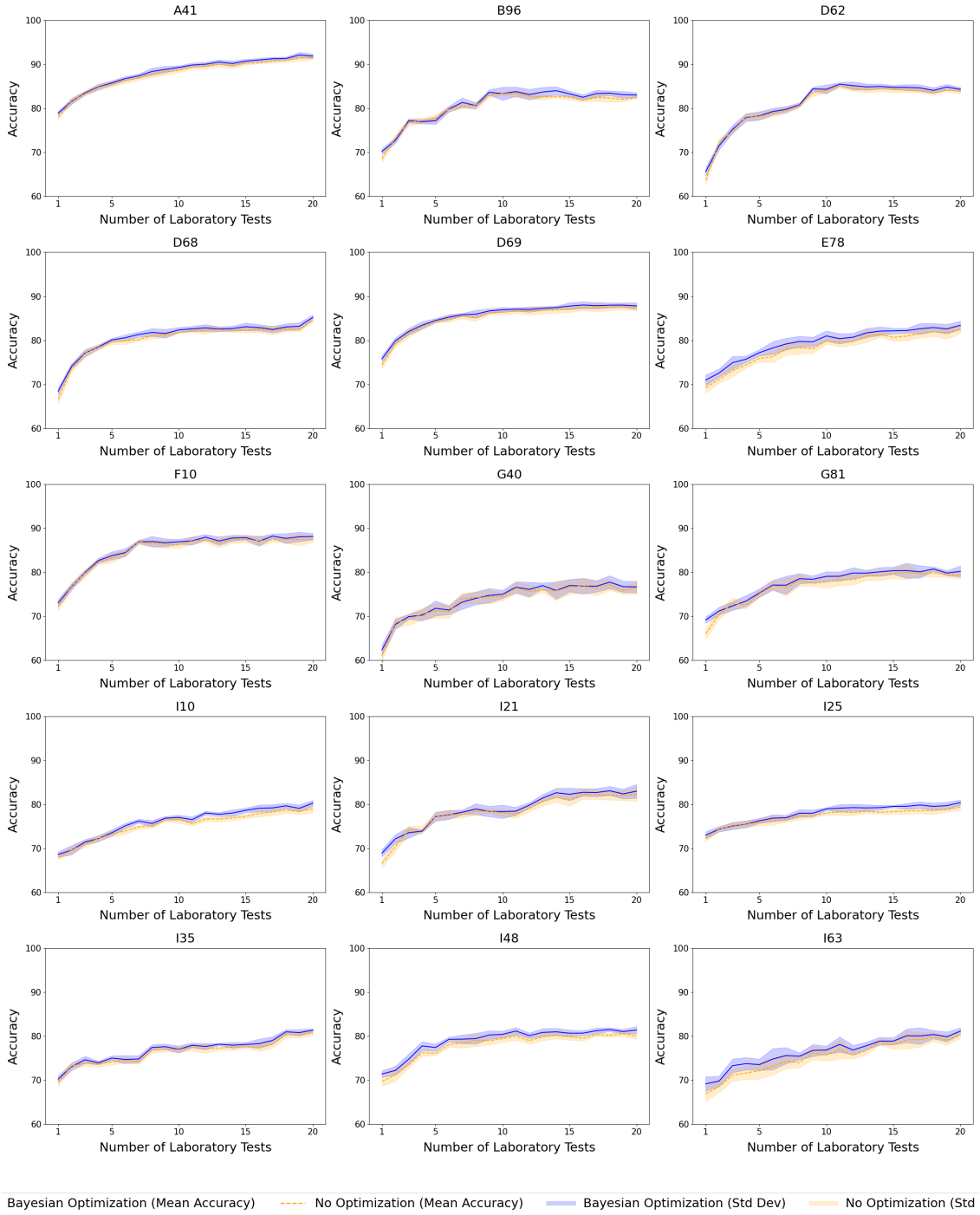

Figure A.3: Line plots comparing the accuracy achieved when aggregating laboratory tests (x-axis: number of laboratory tests, y-axis: accuracy) for 15 diagnoses, with and without hyperparameter tuning using Bayesian optimization. Each line represents the mean accuracy for a diagnosis, with shaded areas indicating the standard deviation.

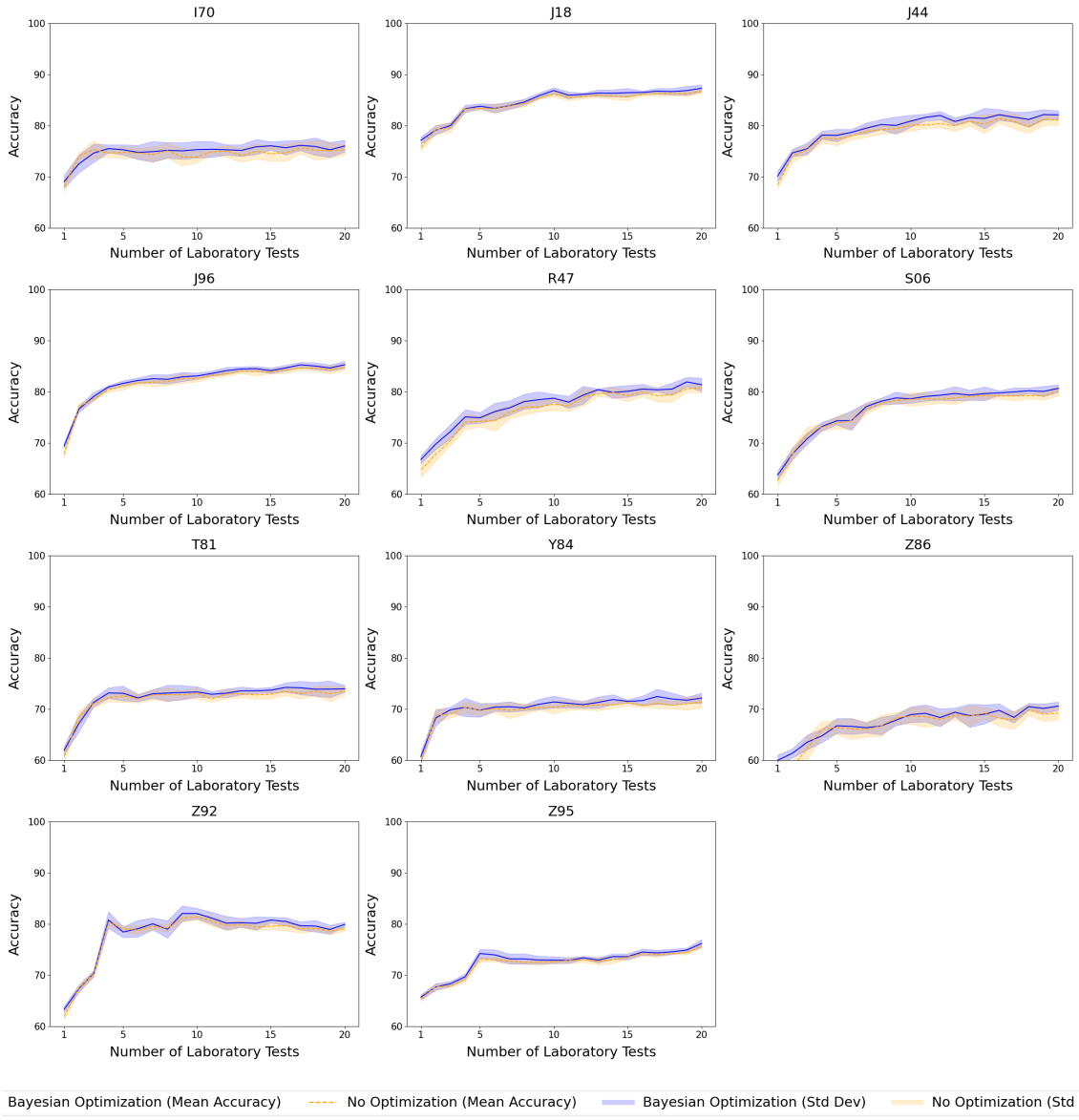

Figure A.4: Line plots comparing the accuracy achieved when aggregating laboratory tests (x-axis: number of laboratory tests, y-axis: accuracy) for 11 diagnoses, with and without hyperparameter tuning using Bayesian optimization. Each line represents the mean accuracy for a diagnosis, with shaded areas indicating the standard deviation.

Appendix B

Supplementary Tables

LOINC	Description
2823-3	Potassium [Moles/volume] in Serum or Plasma
2951-2	Sodium [Moles/volume] in Serum or Plasma
5894-1	Prothrombin time (PT) actual/Normal
1988-5	C reactive protein [Mass/volume] in Serum or Plasma
14749-6	Glucose [Moles/volume] in Serum or Plasma
14682-9	Creatinine [Moles/volume] in Serum or Plasma
6690-2	Leukocytes [#/volume] in Blood by Automated count
777-3	Platelets [#/volume] in Blood by Automated count
718-7	Hemoglobin [Mass/volume] in Blood
789-8	Erythrocytes [#/volume] in Blood by Automated count
4544-3	Hematocrit [Volume Fraction] of Blood by Automated count
787-2	MCV [Entitic volume] by Automated count
785-6	MCH [Entitic mass] by Automated count
786-4	MCHC [Mass/volume] by Automated count
6301-6	INR in Platelet poor plasma by Coagulation assay
788-0	Erythrocyte distribution width [Ratio] by Automated count
32623-1	Platelet mean volume [Entitic volume] in Blood by Automated count
62238-1	Glomerular filtration rate/1.73 sq M.predicted [Volume Rate/Area] in Serum, Plasma or Blood by Creatinine-based formula (CKD-EPI)
5902-2	Prothrombin time (PT)
1743-4	Alanine aminotransferase [Enzymatic activity/volume] in Serum or Plasma by With P-5'-P
30239-8	Aspartate aminotransferase [Enzymatic activity/volume] in Serum or Plasma by With P-5'-P
22664-7	Urea [Moles/volume] in Serum or Plasma
11558-4	pH of Blood
11557-6	Carbon dioxide [Partial pressure] in Blood
11556-8	Oxygen [Partial pressure] in Blood
2075-0	Chloride [Moles/volume] in Serum or Plasma
14979-9	aPTT in Platelet poor plasma by Coagulation assay
20564-1	Oxygen saturation in Blood
59826-8	Creatinine [Moles/volume] in Blood
14927-8	Triglyceride [Moles/volume] in Serum or Plasma
14647-2	Cholesterol [Moles/volume] in Serum or Plasma
14646-4	Cholesterol in HDL [Moles/volume] in Serum or Plasma
39469-2	Cholesterol in LDL [Moles/volume] in Serum or Plasma by calculation
59261-8	Hemoglobin A1c/Hemoglobin.total in Blood by IFCC protocol
46418-0	INR in Capillary blood by Coagulation assay
83071-1	25-Hydroxyvitamin D2+25-Hydroxyvitamin D3 [Moles/volume] in Serum or Plasma by Immunoassay
69419-0	Cholesterol in LDL [Moles/volume] in Serum or Plasma by Direct assay
68438-1	25-Hydroxyvitamin D3+25-Hydroxyvitamin D2 [Moles/volume] in Serum or Plasma
20448-7	Insulin [Units/volume] in Serum or Plasma

Table B.1: All 39 laboratory tests contained within the Swiss BioRef dataset. The tests are sorted from the most common to the least common in the dataset.

Feature	Description
SubjectPseudoIdentifier	Patient identifier
AdministrativeCase	Case identifier
DataProviderInstitute	Data provider identifier
Labtest	Description of the laboratory test
LOINC	LOINC code corresponding to the test
LabResultValue	Measured value
LabResultUnit	Unit of measurement for the corresponding value
Diag01	ICD-10-GM code of relevant diagnosis 1
Diag02	ICD-10-GM code of relevant diagnosis 2
Diag03	ICD-10-GM code of relevant diagnosis 3
Diag04	ICD-10-GM code of relevant diagnosis 4
Diag05	ICD-10-GM code of relevant diagnosis 5
Age	Patient's Age
AgeUnit	Age unit
AgeType	LOINC identifying the age type
AdministrativeGender	Patient's administrative gender
device_udi	Unique device identifier from the Global Unique Device Identification Database
testkit_udi	Device type identifiers from the Global Medical Device Nomenclature
testkit_type	Unique test kit identifier from the Global Unique Device Identification Database
RFIKey_device	Test kit type identifiers from the Global Medical Device Nomenclature

Table B.2: Complete list of features of the Swiss BioRef dataset.

Diagnosis	Number of Cases	Frequency(%)	Description
I25	962,055	16.12	Chronic ischemic heart disease
I10	727,065	12.18	Essential (primary) hypertension
N18	516,478	8.65	Chronic kidney disease (CKD)
I50	498,619	8.36	Heart failure
Z95	442,007	7.41	Presence of cardiac and vascular implants and grafts
I48	396,243	6.64	Atrial fibrillation and flutter
I11	332,533	5.57	Hypertensive heart disease
N17	328,894	5.51	Acute kidney failure
E78	315,122	5.28	Disorders of lipoprotein metabolism and other lipidemias
E11	307,736	5.16	Type 2 diabetes mellitus
Y84	255,273	4.28	Medical procedure causing abnormal reaction or later complication
I63	255,180	4.28	Cerebral infarction (stroke)
X59	244,867	4.1	Exposure to unspecified factor causing injury
E87	225,363	3.78	Other disorders of fluid, electrolyte, and acid-base balance
S06	225,355	3.78	Intracranial injury
I21	211,929	3.55	Acute myocardial infarction (heart attack)
Z92	211,504	3.54	Personal history of medical treatment
I70	207,207	3.47	Atherosclerosis
T81	202,664	3.4	Complications of procedures, not elsewhere classified
R47	200,260	3.36	Speech disturbances
I35	198,955	3.33	Nonrheumatic aortic valve disorders
J96	188,906	3.17	Respiratory failure, not elsewhere classified
D62	187,967	3.15	Acute posthemorrhagic anemia
Z86	176,765	2.96	Personal history of certain other diseases
G40	166,949	2.8	Epilepsy
O09	165,229	2.77	Supervision of high-risk pregnancy
G81	164,962	2.76	Hemiplegia and hemiparesis
A41	163,834	2.75	Sepsis
C77	145,894	2.44	Secondary and unspecified malignant neoplasm of lymph nodes
C78	144,179	2.42	Secondary malignant neoplasm of respiratory and digestive organs
N39	142,302	2.38	Other disorders of urinary system
B96	141,788	2.38	Other bacterial agents as the cause of diseases
D68	137,795	2.31	Other coagulation defects
J44	130,080	2.18	Chronic obstructive pulmonary disease (COPD)
C79	129,548	2.17	Secondary malignant neoplasm of other sites
U99	127,101	2.13	Medical surveillance and observation cases
I34	126,771	2.12	Nonrheumatic mitral valve disorders
D69	124,446	2.09	Purpura and other hemorrhagic conditions
J18	123,277	2.07	Pneumonia, organism unspecified
S02	120,852	2.03	Fracture of skull and facial bones

Table B.3: Most common Diagnoses in the Swiss BioRef dataset.

Hyperparameter	Description	Default Value
n_estimators	Number of decision trees.	100
max_depth	Maximum depth of a tree; controls model complexity.	6
learning_rate	Step size shrinkage; balances weight updates between iterations.	0.3
subsample	Fraction of samples used for training each tree; prevents overfitting.	1.0
colsample_bytree	Fraction of features used for each tree.	1.0
min_child_weight	Minimum sum of weights required for child nodes; prevents overfitting.	1
gamma	Minimum loss reduction required for a split; regularizes the tree.	0
alpha	L1 regularization term; adds sparsity to the model.	0
lambda	L2 regularization term; prevents overfitting.	1
scale_pos_weight	Balances positive and negative classes for imbalanced datasets.	1
objective	Specifies the learning task (e.g., binary:logistic for binary classification).	reg:squarederror
tree_method	Algorithm for constructing trees (e.g., hist for large datasets).	auto

Table B.4: Default values of hyperparameters for the XGBoost classifier using gbtree. Default values are based on the official XGBoost documentation (version 2.1.3).

Bibliography

[1] Longbing Cao. Ai in finance: challenges, techniques, and opportunities. *ACM Computing Surveys (CSUR)*, 55(3):1–38, 2022.

[2] Giuseppe Amato, Malte Behrmann, Frédéric Bimbot, Baptiste Caramiaux, Fabrizio Falchi, Ander García, Joost Geurts, Jaume Gibert, Guillaume Gravier, Hadmut Holken, Hartmut Koenitz, Sylvain Lefebvre, Antoine Liutkus, Fabien Lotte, Andrew Perkis, Rafael Redondo, Enrico Turrin, Thierry Viéville, and Emmanuel Vincent. AI in the media and creative industries. *CoRR*, abs/1905.04175, 2019.

[3] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. Ai in health and medicine. *Nature Medicine*, 28(1):31–38, Jan 2022.

[4] Hafsa Habehh and Suril Gohel. Machine learning in healthcare. *Curr Genomics*, 22(4):291–300, December 2021.

[5] Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics Decis. Mak.*, 19(1):281, 2019.

[6] Lulu Wang. Early diagnosis of breast cancer. *Sensors*, 17(7), 2017.

[7] David S. Celermajer, Clara K. Chow, Eloi Marijon, Nicholas M. Anstey, and Kam S. Woo. Cardiovascular disease in the developing world. *Journal of the American College of Cardiology*, 60(14):1207–1216, 2012.

[8] J. Hjelmesæth, A. Hartmann, T. Leivestad, H. Holdaas, S. Sagedal, M. Olstad, and T. Jenssen. The impact of early-diagnosed new-onset post-transplantation diabetes mellitus on survival and major cardiac events. *Kidney International*, 69(3):588–595, 2006.

[9] Rohan Bhardwaj, Ankita R. Nambiar, and Debojyoti Dutta. A study of machine learning in healthcare. In *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 236–241, 2017.

[10] Vivek Kaul, Sarah Enslin, and Seth A. Gross. History of artificial intelligence in medicine. *Gastrointestinal Endoscopy*, 92(4):807–812, 2020.

[11] Emefa Surprize Deborah Buaka and Md Zubab Ibne Moid. Ai and medical imaging technology: evolution, impacts, and economic insights. *The Journal of Technology Transfer*, 49(6):2260–2272, Dec 2024.

[12] Sarah B. Scruggs, Karol Watson, Andrew I. Su, Henning Hermjakob, John R. Yates, Merry L. Lindsey, and Peipei Ping. Harnessing the heart of big data. *Circulation Research*, 116(7):1115–1119, 2015.

[13] Sarah J. MacEachern and Nils D. Forkert. Machine learning for precision medicine. *Genome*, 64(4):416–425, 2021. PMID: 33091314.

[14] Naveed Rabbani, Grace Y.E. Kim, Carlos J. Suarez, and Jonathan H. Chen. Applications of machine learning in routine laboratory medicine: Current state and future directions. *Clinical Biochemistry*, 103:1–7, 2022.

[15] Yoon Hyung-Jin Lee Choong Ho. Medical big data: promise and challenges. *Kidney Res Clin Pract*, 36(1):3–11, 2017.

[16] Tobias Ueli Blatter, Harald Witte, Jules Fasquelle-Lopez, Christos Theodoros Nakas, Jean Louis Raisaro, and Alexander Benedikt Leichtle. The bioref infrastructure, a framework for real-time, federated, privacy-preserving, and personalized reference intervals: Design, development, and application. *Journal of Medical Internet Research*, 25:e47254, 2023.

[17] Ziad Obermeyer and Ezekiel J. Emanuel. Predicting the future — big data, machine learning, and clinical medicine. *New England Journal of Medicine*, 375(13):1216–1219, 2016.

[18] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[19] Jenna Wong, Mara Murray Horwitz, Li Zhou, and Sengwee Toh. Using machine learning to identify health outcomes from electronic health record data. *Current Epidemiology Reports*, 5(4):331–342, Dec 2018.

[20] Joshua C. Denny. Chapter 13: Mining electronic health records in the genomics era. *PLOS Computational Biology*, 8(12):1–15, 12 2012.

[21] Suraj Rajendran, Zhenxing Xu, Weishen Pan, Arnab Ghosh, and Fei Wang. Data heterogeneity in federated learning with electronic health records: Case

studies of risk prediction for acute kidney injury and sepsis diseases in critical care. *PLOS Digital Health*, 2(3):1–26, 03 2023.

[22] Effy Vayena, Alessandro Blasimme, and I. Glenn Cohen. Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine*, 15(11):1–4, 11 2018.

[23] M Mostafizur Rahman and Darryl N Davis. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, 3(2):224, 2013.

[24] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[25] Gregor Gunčar, Matjaž Kukar, Mateja Notar, Miran Brvar, Peter Černelč, Manca Notar, and Marko Notar. An application of machine learning to haematological diagnosis. *Scientific Reports*, 8(1):411, Jan 2018.

[26] Dong Jin Park, Min Woo Park, Homin Lee, Young-Jin Kim, Yeongsic Kim, and Young Hoon Park. Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Scientific Reports*, 11(1):7567, Apr 2021.

[27] Fabian Gribi. Binary classification of blood values. *Bachelor Thesis, University of Bern*, 2023.

[28] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[29] Wei Li, Yanbin Yin, Xiongwen Quan, and Han Zhang. Gene expression value prediction based on xgboost algorithm. *Frontiers in genetics*, 10:1077, 2019.

[30] Adeola Ogunleye and Qing-Guo Wang. Xgboost model for chronic kidney disease diagnosis. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(6):2131–2140, 2019.

[31] Junhui Wang and Michael Gribskov. Irespy: an xgboost model for prediction of internal ribosome entry sites. *BMC bioinformatics*, 20(1):1–15, 2019.

[32] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[33] Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, Dec 1998.

[34] Clement J McDonald, Stanley M Huff, Jeffrey G Suico, Gilbert Hill, Dennis Leavelle, Raymond Aller, Arden Forrey, Kathy Mercer, Georges DeMoor, John Hook, Warren Williams, James Case, Pat Maloney, and for the Laboratory LOINC Developers. Loinc, a universal standard for identifying laboratory observations: A 5-year update. *Clinical Chemistry*, 49(4):624–633, 04 2003.

[35] Bernd Graubner. *ICD-10-GM 2014 Systematisches Verzeichnis: Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme 11. Revision-German Modification Version 2014*. Deutscher Ärzteverlag, 2013.