Fachbereich Mathematik und Informatik
Machine Learning for Medical Data

Abschlussarbeit im Bachelorstudiengang Informatik

Bachelorarbeit von

Iraj Masoudian

Matr.-Nr. 3041956

Using Machine Learning for Non-Invasive Disease Prediction to Derive Causal Discovery

Winter Semester 2024/2025
zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.)

Erstgutachter:	Prof. Dr. Dominik Heider
Zweitgutachter:	Prof. Dr. Gunnar Klau
Abgabedatum:	13.11.2024

Selbstständigkeitserklärung

Hiermit versichere ich, **Iraj Masoudian**, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Ausführungen, die fremden Quellen wörtlich oder sinngemäß entnommen wurden, sind kenntlich gemacht.

Düsseldorf, den 13.11.2024

Iraj Masoudian

Abstract

This thesis addresses the need for accessible, non-invasive diagnostic tools in healthcare, driven by the global shortage of medical resources. Focusing on diseases with high prevalence and mortality, this thesis uses machine learning to develop a predictive model framework for eight conditions such as heart disease and breast cancer. By using non-invasive data, we aim to reduce patient risk, the need for cost intensive procedures and increase diagnostic accessibility.

The implementation involves training discriminative models, including decision trees, random forests, and neural networks. Feature importance analyses were conducted, and the models were assessed for both their full feature sets and non-invasive subsets to evaluate predictive power and generalization. Intervention analyses with synthetic data and pivot tables were used to assess feature sensitivity and model generalization which can't be observed on the test set if the dataset has an inherent bias. For further intervention analysis a causal discovery was conducted to see whether a feature had a causal relationship to the respective condition to derive active intervention possibilities.

Results indicate that non-invasive models achieved comparable predictive accuracy to models trained on all features for several diseases, with some limitations in generalization due to inherent biases rather than feature reduction. The pivot tables and calibration method highlighted key insights, such as the influence of specific features on disease predictions, revealing the medical relevance of certain features. While non-invasive models alone were not universally superior but also not worse than the models trained on all features, they offer meaningful predictions that could guide further diagnostic testing, including invasive procedures. The intervention analysis and causal discovery showed the limitation of deriving active interventions based on a discriminative model, highlighting the need for generative models that learn the data distribution instead of a decision boundary.

In conclusion, this thesis presents an analytical framework for deriving predictive discriminative models from non-invasive data, contributing to the development of accessible, non-invasive diagnostic tools and offering insights into the potential of machine learning in clinical decision support.

Zusammenfassung

Diese Arbeit befasst sich mit dem Bedarf an zugänglichen, nicht-invasiven Diagnosewerkzeugen im Gesundheitswesen, der sich aus der weltweiten Knappheit medizinischer Ressourcen ergibt. Die Arbeit konzentriert sich auf Krankheiten mit hoher Prävalenz und Sterblichkeit und verwendet maschinelles Lernen, um Vorhersagemodelle für acht Krankheiten wie Herzerkrankungen und Brustkrebs zu entwickeln. Durch die Nutzung nicht-invasiver Daten wollen wir das Patientenrisiko und den Bedarf an kostenintensiven Verfahren reduzieren und die diagnostische Zugänglichkeit verbessern.

Die Implementierung umfasst das Training von Diskriminanzmodellen, einschließlich Decision Tree, Random Forest und neuronalen Netzen. Es wurden Signifikanzanalysen von Merkmalen durchgeführt und die Modelle sowohl für ihre vollständigen Merkmalssätze als auch für nichtinvasive Merkmale evaluiert, um ihre Vorhersagekraft und Verallgemeinerbarkeit zu beurteilen. Interventionsanalysen mit synthetischen Daten und Pivot-Tabellen wurden verwendet, um die Sensitivität der Merkmale und die Generalisierung der Modelle zu bewerten, die bei inhärenter Verzerrung des Datensatzes in den Testdatensätzen nicht beobachtet werden konnten. Für weitere Interventionsanalysen wurde eine kausale Analyse durchgeführt, um zu sehen, ob ein Merkmal eine kausale Beziehung zu einer bestimmten Krankheit hat, um aktive Interventionsmöglichkeiten abzuleiten.

Die Ergebnisse zeigen, dass nicht-invasive Modelle für mehrere Krankheiten eine vergleichbare Vorhersagegenauigkeit erreichen wie Modelle, die auf allen Merkmalen trainiert wurden, mit einigen Einschränkungen bei der Generalisierbarkeit, die eher auf inhärente Verzerrungen als auf eine Reduzierung der Merkmale zurückzuführen sind. Die Pivot-Tabellen und die Kalibrierungsmethode haben wichtige Erkenntnisse hervorgebracht, wie z. B. den Einfluss spezifischer Merkmale auf die Krankheitsvorhersage, was die medizinische Relevanz bestimmter Merkmale unterstreicht. Obwohl die nicht-invasiven Modelle allein nicht durchweg besser oder schlechter waren als die auf allen Merkmalen trainierten Modelle, lieferten sie aussagekräftige Vorhersagen, die weitere diagnostische Tests, einschließlich invasiver Verfahren, leiten können. Die Interventionsanalyse und die Kausalitätserkennung zeigten die Grenzen der Ableitung aktiver Interventionen auf der Grundlage eines diskriminierenden Modells auf und unterstreichen die Notwendigkeit generativer Modelle, die die Verteilung der Daten und nicht eine Entscheidungsgrenze lernen.

Zusammenfassend stellt diese Arbeit einen analytischen Rahmen für die Ableitung prädiktiver diskriminierender Modelle aus nicht-invasiven Daten vor, der zur Entwicklung zugänglicher nicht-invasiver Diagnosewerkzeuge beiträgt und Einblicke in das Potenzial des maschinellen Lernens für die klinische Entscheidungsunterstützung bietet.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Introduction

The healthcare industry faces a number of challenges, including workforce shortages, bad care in rural areas, and poor working conditions for healthcare professionals [1, 2, 3]. These issues not only strain existing resources but also impacts access to quality care [4]. To address these challenges, there is a growing interest in advancing technologies, for instance machine learning, to enhance healthcare and decision-making [5].

Machine learning is a branch of artificial intelligence that enables computers to analyze data and make informed predictions or decisions based on that data, rather than relying on rules or heuristics. By identifying patterns in datasets, machine learning algorithms can uncover insights that might not be easily noticed by human experts. In the context of healthcare, machine learning serves as a transformative tool, offering new approaches to analyze patient data and generate accurate predictions [6, 7].

The application of machine learning in healthcare has shown promising results in various areas, for instance in disease diagnosis, treatment planning and predictive analytics [8]. By extracting meaningful insights from vast amounts of medical data, machine learning aids healthcare professionals in making informed decisions, ultimately enhancing clinical decision-making and diagnostic accuracy [9].

One attempt to advance AI-driven healthcare is the Virtual Doctor [10]. Designed to interact with patients using speech recognition and synthesis. The Virtual Doctor differs from traditional AI systems by engaging directly with patients. It collects anamnesis and utilizes non-invasive sensors to predict diseases, such as Type 2 Diabetes Mellitus (T2DM). This system tries to address shortage of healthcare providers, particularly in rural areas, by providing primary medical assessments and suggesting further diagnostic steps based on features like BMI, age, and lifestyle factors.
This proof-of-concept is part of the Virtual Doctor system, developed for T2DM prediction, which combines patient interviews with biometric data to generate a risk estimation, recommending diagnostic tests like the HbA1c blood test when necessary. The system not only aids in early detection but also relieves strain on healthcare resources by autonomously handling routine diagnoses and offering recommendations for further medical care. This concept highlights the growing role of AI in enhancing healthcare access and efficiency, especially where traditional medical resources are scarce.

This thesis also examines the significance of invasive and non-invasive features in disease prediction, aiming to develop more use cases for the Virtual Doctor. Invasive features often

involve diagnostic methods that require more intrusive measures, such as blood tests or biopsies. While these tests can provide valuable information and are often considered the standard for diagnosis, their use can lead to increased patient discomfort, and healthcare costs [11, 12]. Additionally, logistical challenges arise, especially in rural areas, where access to invasive testing may be limited [13, 14].

In contrast, non-invasive features are derived from less intrusive procedures, such as demographic data, medical history, or symptoms reported by the patient. The ability to accurately predict diseases using non-invasive methods could significantly enhance patient comfort and reduce costs within healthcare systems. If non-invasive features can demonstrate comparable accuracy to invasive tests, healthcare systems could reduce their reliance on intrusive procedures while also offering recommendations for necessary invasive analysis, thereby improving patient care and accessibility to diagnostic services.

Finally, an essential component of this thesis is exploring how diagnoses derived from machine learning models can inform and guide interventions. By making use of non-invasive features in predictive models, healthcare providers can identify risk patients early, allowing for timely and targeted interventions. The ability to derive interventions from accurate diagnoses can help mitigate healthcare disparities, especially in rural or resource-limited areas, offering essential services that might otherwise be inaccessible due to logistical challenges associated with invasive procedures.

This thesis aims to provide insights into how effective disease prediction and intervention derivation can be achieved using non-invasive methods, while also developing an analytical framework to derive meaningful models based on non-invasive data, ultimately contributing to improved patient care and resource allocation in healthcare systems.

2. Background

2.1. Medical Background

In this thesis several disease datasets were used to derive a comprehensive analysis framework. Understanding the nature of each disease is crucial for understanding the importance of non-invasive prediction methods.

2.1.1. Heart Disease

Heart disease has a range of conditions, for instance coronary artery disease, arrhythmia, and congenital heart defects [15, 16, 17]. It is one of the leading causes of death worldwide, taking an estimated 17.9 million lives each year, where 80% of the occurrences are from low to middle income countries [18, 19]. Early diagnosis and treatment are critical in preventing heart disease [20]. Traditional diagnostic methods such as angiography are invasive procedures that involve the use of catheters [21]. Angiography is a medical imaging technique where a contrast dye is injected into the blood vessels, allowing detailed X-ray images of the coronary arteries [21]. This helps in identifying any abnormalities in the arteries, making it an important tool for diagnosing conditions like coronary artery disease [21]. Non-invasive prediction methods for heart disease include tools such as electrocardiograms and imaging techniques [22]. Electrocardiograms record the electrical activity of the heart and can help identify irregularities that may indicate heart disease and reduce the need for more invasive procedures [23].

2.1.2. Breast Cancer

Breast cancer originates from the breast tissue, usually from the lining of the milk ducts or lobules [24, 25]. Breast cancer is one of the most common cancer types in women, with high incidence rates globally [26]. In 2020, there were over 2.3 million new cases and 685,000 deaths due to breast cancer, making it the most commonly diagnosed cancer worldwide [26]. Most cases occur in developing countries [26]. The future burden of breast cancer is predicted to increase significantly, with projections suggesting over 3 million new cases and 1 million deaths by 2040 [26]. Early detection of breast cancer significantly improves prognosis, with screening through mammography being a standard practice.

Detecting second breast cancers in the asymptomatic phase not only leads to early-stage identification but also improves relative survival rates by 27% to 47% [27, 28]. However, mammography alone has limitations, including the potential for false positives, where non-cancerous findings are mistakenly identified as cancer, and false negatives, where actual cancers go undetected. False positives can lead to unnecessary anxiety, additional imaging, and invasive procedures such as biopsies, which can be both physically and emotionally taxing for patients [29, 30]. Non-invasive machine learning models, when combined with imaging techniques, can enhance early detection, potentially reducing the need for invasive biopsies and improving patient outcomes [31].

2.1.3. Alzheimer's Disease

Alzheimer disease is a disease that leads to memory loss, cognitive decline, and eventually an inability to carry out daily tasks [32, 33]. It is the most common form of dementia, affecting older people [33]. Diagnosing Alzheimer is challenging and often involves invasive procedures like spinal taps and biopsies. A spinal tap, or lumbar puncture, collects cerebrospinal fluid to analyze biomarkers indicative of Alzheimer, such as amyloid-beta and tau proteins [34]. This procedure can cause discomfort and has risks, including headaches and infection [34]. Brain biopsies, which involve removing tissues for examination, aren't done often due to their invasive nature and risk for complications [35]. Non-invasive techniques, for instance brain imaging like magnetic resonance imaging and positron emission tomography or biomarker analysis like the novel gene ALZAS, combined with machine learning algorithms, hold promising results in predicting Alzheimer disease in its early stages and allowing for interventions that could slow its progression [36, 37, 38].

2.1.4. Depression

Depression is a mental health disorder characterized by persistent sadness, lack of interest in activities, and a variety of emotional and physical problems [39]. It affects how a person feels, thinks, and handles daily activities [40]. Traditional diagnostic methods rely heavily on self-reported symptoms and clinician observations [41]. Globally, an estimated 5% of adults experience depression, with women being affected more frequently than men [42]. However, recent advancements in machine learning allow for the identification of depression based on non-invasive data sources, such as speech patterns, facial expressions, and physiological measurements [43, 44, 45]. Physiological measurements can include heart rate variability, and skin temperature, which often correlate with emotional states and can provide subtle indicators of mental health status [45]. These methods can provide a more objective and earlier diagnosis [46].

2.1.5. Chronic Kidney Disease

Chronic kidney disease (CKD) refers to the gradual loss of kidney function over time. CKD is a progressive condition that affects more than 10% of the population worldwide [47]. As the kidney functionality worsens, they are less able to filter waste and excess fluids from the blood, which can lead to end-stage kidney disease [48]. Diagnosing CKD typically involves invasive procedures such as biopsies, where a small sample of kidney or bone tissue is taken and examined to assess damage or abnormalities [49]. In bone biopsies, often used for patients with CKD, a piece of bone (usually from the hip) is analyzed to monitor mineral imbalances affecting bone density and structure, providing insight into CKD-associated bone disorders [49]. However, machine learning techniques using non-invasive biomarkers, like urine tests, have the potential to detect CKD in its early stages, thereby preventing further damage and improving patient outcomes [50].

2.1.6. Multiple Sclerosis

Multiple sclerosis (MS) is a chronic autoimmune disease in which the immune system attacks the protective covering of nerves, leading to communication problems between the brain and the rest of the body [51]. This can result in physical and cognitive impairments [51]. The estimated number of people with MS worldwide has increased to 2.8 million in 2020 [52]. MS diagnosis often involves invasive procedures like spinal taps or biopsies [53]. In a spinal tap (or lumbar puncture), a needle is inserted into the lower back to extract cerebrospinal fluid, which is then analyzed for signs of inflammation or immune proteins indicative of MS [54]. This helps to rule out other neurological conditions and confirm an MS diagnosis [54]. Biopsies, though less common, may involve extracting a small tissue sample when there is ambiguity in diagnosis, often to differentiate MS from other similar neurological diseases [55]. However, non-invasive tools, such as MRI scans, when combined with machine learning algorithms, can aid in predicting the onset and progression of MS, allowing for earlier intervention [56]. MRI (Magnetic Resonance Imaging) uses powerful magnets and radio waves to create detailed images of the brain and spinal cord, which are crucial for detecting MS-related lesions [57]. These scans allow physicians to visualize changes in brain tissue, spot inflammation, and monitor the formation of new lesions over time, all without exposing the patient to radiation [56, 57].

2.1.7. Differentiated Thyroid Cancer Recurrence

Differentiated thyroid cancer (DTC) arises from the follicular or papillary cells of the thyroid gland and generally has a good prognosis when treated early [58]. However, recurrence is possible, making long-term monitoring essential [59]. DTC recurrence is not uncommon, occurring in 3–13% of low-risk patients, 21–36% of intermediate-risk patients, and approximately 68% of high-risk patients [60]. Current methods for detecting

differentiated thyroid cancer recurrence often rely on imaging and invasive biopsies [58, 59]. Imaging techniques, such as ultrasound, are frequently used to detect abnormal growths or tissue changes that could indicate a recurrence [61]. Ultrasound, the most common imaging tool, provides detailed images of the thyroid and surrounding tissues to identify suspicious nodules or lymph nodes [62]. Invasive biopsies, often fine-needle aspiration (FNA) biopsies, are performed when imaging reveals suspicious nodules [63]. In this procedure, a thin needle is used to collect tissue or fluid from the thyroid or surrounding lymph nodes for laboratory analysis [64]. This allows for a more definitive diagnosis but can be uncomfortable for patients and carries a small risk of complications, such as bleeding or infection [63, 64]. Machine learning models applied to non-invasive data, such as biochemical markers and ultrasound imaging, can help predict the likelihood of recurrence, reducing the need for invasive follow-up procedures [65]. An example of a biochemical marker is thyroglobulin, a protein produced by thyroid cells [66]. Elevated levels of thyroglobulin in blood tests can indicate the presence of recurrent thyroid cancer [66].

2.1.8. Pathological Voice Disorders

Pathological voice disorders occur due to damage to the vocal cords or related structures, often leading to changes in speech quality [67]. These disorders are typically diagnosed using invasive techniques such as laryngoscopy [68]. Laryngoscopy involves the use of a thin, flexible tube equipped with a camera and light source, known as a laryngoscope [68]. This instrument is inserted through the mouth or nose to directly visualize the vocal cords and surrounding structures of the larynx [68]. The procedure allows healthcare providers to assess any abnormalities, such as inflammation, lesions, or nodules on the vocal cords, which can contribute to voice disorders [69]. While it provides valuable diagnostic information, laryngoscopy can cause discomfort and may require local anesthesia for patient comfort [69]. Non-invasive alternatives, such as acoustic analysis of voice recordings, can be used to detect abnormalities in vocal patterns [70, 71]. Machine learning models can analyze voice data to identify pathological conditions, offering a less invasive and more accessible diagnostic tool for detecting voice disorders early [72, 73, 74].

2.1.9. Diabetes

Diabetes Mellitus (DM) is a chronic disease where the body struggles to regulate blood sugar levels, typically due to insulin production issues or insulin resistance [75]. There are two main types [75]. One of them is type 1 Diabetes [76]. Type 1 Diabetes is an autoimmune condition often diagnosed early in life, where the immune system destroys insulin-producing cells, necessitating lifelong insulin therapy [76]. Type 2 Diabetes is more common, typically develops in adults, and results from the body's reduced sensitivity to insulin or a decline in insulin production, often linked to lifestyle factors [77].

Diabetes poses significant health risks, including heart disease, kidney damage, and vision loss, and affects over 500 million people worldwide [78, 79, 80, 81]. Non-invasive diagnostic methods using machine learning are gaining attention for their potential to improve early detection and intervention, especially in rural areas [10, 82].

2.1.10. Invasive vs. Non-Invasive Features

In healthcare, features can be broadly categorized into invasive and non-invasive. Invasive features refer to data obtained through procedures that penetrate the body, such as blood tests, or biopsies [83]. These procedures, while informative, can be uncomfortable for patients, expensive, and resource-intensive [83]. Non-invasive features, on the other hand, include data such as demographic information, for instance age or gender, and self-reported symptoms. These features are easier to collect and pose minimal risk or discomfort to patients [83].

2.2. Machine Learning Background

In this section, we will delve into the machine learning techniques that are applicable for disease classification. Each method has its own strengths and weaknesses, and understanding these will provide insights into their applicability for non-invasive disease prediction.

2.2.1. Support Vector Machine (SVM)

Support Vector Machines (SVMs) are supervised learning models used for classification and regression tasks [84]. The basic idea behind SVM is to find a decision boundary, called a hyperplane, that separates the data into different classes. SVM aims to maximize the margin between the hyperplane and the closest data points from each class. These closest data points are known as support vectors, and they are important in determining the position of the hyperplane. For linearly separable data, the goal of SVM is to find a hyperplane in an n-dimensional space (where n is the number of features) that separates the data points into different classes. The hyperplane is defined by the equation:

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{2.1}$$

where $\mathbf{w}$ is the weight vector (normal to the hyperplane), $\mathbf{x}$ is the feature vector and b is the bias term.

The decision boundary is defined by the maximized distance (margin) between the hyperplane and the nearest data points from each class. The margin is defined as $\frac{1}{||\mathbf{w}||}$, and the optimization problem is to minimize $||\mathbf{w}||$ while ensuring that the data points are correctly

classified:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad \forall i \tag{2.2}$$

Here, $y_i \in \{-1, 1\}$ represents the class labels.

In many cases, the data is not linearly separable in the original feature space. To handle such cases, SVM uses the kernel trick, which implicitly maps the input data into a higher-dimensional feature space where it becomes linearly separable. This transformation is done through a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$, which computes the inner product of the data points in the new space without explicitly performing the transformation.

A common kernel is the polynomial kernel, which is defined by:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + c)^d \tag{2.3}$$

where c and d are the kernel parameters.

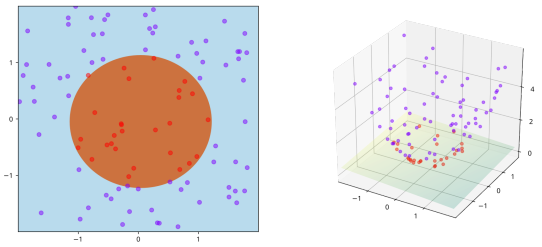


Abbildung 2.1.: This figure shows a training example of SVM where on the left, the data can be seen in two dimensional form, and on the right, the kernel trick applied with a kernel given by $\phi((a, b)) = (a, b, a^2 + b^2)$, which makes it linearly separable in a higher-dimensional feature space. The image was taken from [85].

The optimization problem in SVM can be formulated as a convex optimization problem, which can be solved efficiently. The dual form of the SVM optimization problem is:

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{2.4}$$

subject to:

$$0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{2.5}$$

where α_i are the Lagrange multipliers, and C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing classification error.

2.2.2. k-Nearest Neighbors (KNN)

The k-Nearest Neighbors (KNN) algorithm is a simple, non-parametric model used for classification and regression tasks [86]. It works by making predictions based on the proximity of data points in a feature space.

KNN does not have an explicit training phase, instead, it stores all the training data and classifies new data points by looking at the labels of the k-nearest neighbors. To classify a new data point, the algorithm computes the distance between the new point and all the training points using a distance metric, typically the Euclidean distance, given by:

$$d(\mathbf{x}, \mathbf{x}_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{ij})^2} \tag{2.6}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the new data point, and $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ represents a training data point. The algorithm selects the k closest training points based on this distance.

Once the k-nearest neighbors are identified, the algorithm looks at their corresponding labels. The predicted class for the new data point is the most common class (majority vote) among these k-nearest neighbors. In mathematical terms, the predicted class $\hat{y}$ can be expressed as:

$$\hat{y} = \text{argmax}_c \sum_{i \in \mathcal{N}_k} \mathbb{I}(y_i = c) \tag{2.7}$$

where $\mathcal{N}_k$ is the set of the k-nearest neighbors, y_i is the label of the i-th neighbor, and $\mathbb{I}$ is the indicator function that equals 1 if $y_i = c$, and 0 otherwise.

KNN works well when the data is well-distributed and the decision boundary is complex, but it has some limitations.

In high-dimensional spaces, the distance between points becomes less meaningful due to the sparsity of data. As the number of dimensions increases, the volume of the space grows exponentially, and data points become more spread out. This causes the distances between neighbors to become similar, making it difficult to distinguish between relevant and irrelevant points. This phenomenon is known as the curse of dimensionality.

Since KNN requires storing the entire training dataset and computing distances for every new data point, it can become computationally expensive, especially for large datasets.

In a standard KNN setup, all the training data is stored and used for predictions, which may raise privacy concerns. This is particularly important when dealing with sensitive data such as medical records.

2.2.3. Decision Trees and Random Forests

2.2.3.1. Decision Trees

Decision Trees are a popular supervised learning algorithm used for classification and regression tasks [87]. They work by recursively splitting the data into subsets based on feature values, creating a tree-like structure. Each internal node of the tree represents a decision based on the value of a feature, and each leaf node represents a final output (class label for classification or a continuous value for regression).

The construction of a decision tree involves selecting the best feature to split the data at each node. Common criteria for making these decisions include the gini impurity. The Gini impurity measures the impurity of a node, with lower values indicating a purer node. It is calculated as:

$$Gini(D) = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{2.8}$$

where p_i is the proportion of instances of class i in dataset D and C is the number of classes.

Another criteria is the information gain. Information gain measures the reduction in entropy after a dataset is split on a feature. The entropy $H(D)$ of a dataset D is defined as:

$$H(D) = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{2.9}$$

Information gain is calculated as:

$$IG(D, A) = H(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} H(D_v) \tag{2.10}$$

where D_v is the subset of data for feature A with value v.

Once the tree is constructed, it can make predictions by traversing from the root to a leaf node based on the input features.

2.2.3.2. Random Forest

Random Forest is an ensemble learning model that builds multiple decision trees and aggregates their results to make a prediction [88]. Each tree in the forest is trained on a random subset of the data, which is called bootstrapping, and at each split in the tree, a random subset of features is considered. This randomness reduces the chance of overfitting

and ensures that each tree is less correlated with the others, improving the overall models generalization ability.

In classification problems, Random Forest predicts the class by taking a majority vote from all the individual decision trees:

$$\hat{y} = \text{argmax}_c \sum_{j=1}^{T} \mathbb{I}(y_j = c) \tag{2.11}$$

where T is the total number of trees in the forest, y_j is the predicted class of tree j, and $\mathbb{I}$ is the indicator function that equals 1 if $y_j = c$.

Random Forest is highly effective for both classification and regression tasks and is particularly valued for its ability to handle high-dimensional datasets and provide feature importance metrics, which can help identify the most influential features in the model.

2.2.3.3. XGBoost

XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable implementation of the gradient boosting algorithm, which builds an ensemble of weak learners (typically decision trees) in a sequential manner [89]. Each tree is trained to correct the errors made by the previous ones. The key idea behind XGBoost is to minimize a loss function by adding new models that reduce the residual errors of the previous models. What sets XGBoost apart is its use of regularization techniques (such as L1 and L2 penalties) to prevent overfitting, as well as its ability to handle missing data. XGBoost is known for its high performance and speed, often yielding state-of-the-art results in many machine learning competitions.

2.2.3.4. Gradient Boosting Classifier

Gradient Boosting Classifier is another ensemble method that works by building multiple decision trees sequentially [90]. Unlike Random Forest, where trees are built independently, Gradient Boosting builds each new tree to correct the errors made by the previous trees. The model optimizes a loss function (e.g., binary cross-entropy for classification) by gradually improving its predictions. At each step, the model computes the residuals (the difference between the actual values and the predicted values) and fits a new decision tree to predict these residuals. This process continues until the model reaches a predefined number of trees or the residuals are minimized. While Gradient Boosting is highly accurate, it can be prone to overfitting if not properly regularized, and it may be slower to train than Random Forest due to its sequential nature.

2.2.3.5. Feature Importance in Predictive Models

Feature importance analysis is a critical method used to evaluate the contribution of individual variables in machine learning models [91]. It helps to identify the most influential features that drive the predictions. Models like Random Forests are particularly useful for this analysis, as they naturally provide a ranking of feature importance by assessing the impact of each feature on the model's decision-making process. This insight is valuable for medical research, as it helps to understand which features are most predictive and can guide feature selection in future models.

In medical datasets, invasive features often include measurements or tests that require more intrusive procedures (e.g., blood tests or biopsies), while non-invasive features are typically derived from less intrusive sources, such as demographic data or symptoms reported by the patient.

Random Forest inherently provides a measure of feature importance by calculating how much each feature contributes to the model's predictive power. This allows for an in-depth analysis of the most influential features across the datasets.

2.2.4. Neural Networks

Neural Networks (NNs) are a class of machine learning models inspired by the structure and functioning of the human brain [92]. They are particularly well-suited for modeling complex, non-linear relationships in data and are widely used in areas such as image recognition, natural language processing, and medical diagnostics. A neural network consists of layers of interconnected nodes, also known as neurons, with each connection associated with a weight that adjusts as the network learns from data.

The basic structure of a neural network is composed of three types of layers: the input layer, hidden layers, and the output layer. The input layer serves as the entry point where the network receives the input data, with each node in the input layer corresponding to a feature from the dataset. The hidden layers, which are located between the input and output layers, are where the model learns patterns in the data. Each node in a hidden layer performs a weighted sum of the inputs, applies an activation function (such as ReLU, sigmoid, or tanh) to introduce non-linearity, and passes the result to the next layer [93, 94]. The output layer produces the final predictions of the model. For classification tasks, the output layer often returns a probability distribution (in the case of multi-class classification) or a single probability value (in binary classification) by applying an activation function such as softmax or sigmoid [94].

The training process for neural networks involves several key steps, including forward propagation, loss calculation, backpropagation, and weight updates.

During forward propagation, input data is fed into the network, and the output is calculated layer by layer. The output $a^{(l)}$ of the l-th layer is computed as follows:

$$z^{(l)} = W^{(l)}a^{(l-1)} + b^{(l)} \tag{2.12}$$

$$a^{(l)} = f(z^{(l)}) \tag{2.13}$$

where $W^{(l)}$ are the weights, $b^{(l)}$ are the biases, f is the activation function, and $a^{(l-1)}$ is the output from the previous layer.

After obtaining the predictions, the model computes the loss, which quantifies the difference between the predicted values and the actual target values. For binary classification, the loss can be calculated using binary cross-entropy:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{2.14}$$

where y is the true label, $\hat{y}$ is the predicted probability, and N is the number of samples [95].

In the backpropagation step, the model calculates the gradients of the loss with respect to the weights by applying the chain rule. This process starts from the output layer and propagates backward through the network. The gradient of the loss L with respect to the weights $W^{(l)}$ is given by:

$$\frac{\partial L}{\partial W^{(l)}} = \frac{\partial L}{\partial a^{(l)}} \cdot \frac{\partial a^{(l)}}{\partial z^{(l)}} \cdot \frac{\partial z^{(l)}}{\partial W^{(l)}} \tag{2.15}$$

Once the gradients are calculated, the weights are updated using an optimization algorithm, such as gradient descent. Gradient descent is an iterative optimization algorithm used to minimize the loss function by updating the model's parameters (e.g., weights). The basic idea behind gradient descent is to adjust the model parameters in the direction of the steepest descent (i.e., the negative gradient) of the loss function, which helps the model gradually converge to a minimum.

The update rule for gradient descent can be formulated as follows:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta) \tag{2.16}$$

where:

- θ_t represents the current model parameters (weights) at iteration t,
- η is the learning rate, which controls the step size of the updates,

- $\nabla_\theta L(\theta)$ is the gradient of the loss function with respect to the model parameters,
- θ_{t+1} are the updated parameters after applying the gradient step.

There are several variations of gradient descent, including batch gradient descent, stochastic gradient descent (SGD), and mini-batch gradient descent [96].

Adam (Adaptive Moment Estimation) is a more advanced optimization algorithm that combines the advantages of two popular methods: AdaGrad and RMSProp [97]. Adam adapts the learning rate for each parameter by computing both the first moment (mean) and the second moment (uncentered variance) of the gradients.

Adam maintains two moving averages for each parameter: one for the gradient (first moment) and one for the squared gradient (second moment). These moving averages are updated as follows:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)\nabla_\theta L(\theta_t) \tag{2.17}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)(\nabla_\theta L(\theta_t))^2 \tag{2.18}$$

where:

- m_t is the first moment estimate (mean of gradients),
- v_t is the second moment estimate (variance of gradients),
- β_1 and β_2 are exponential decay rates for the moment estimates (commonly set to 0.9 and 0.999, respectively).

To correct for bias introduced during initialization, the moment estimates are bias-corrected as follows:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{2.19}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{2.20}$$

Finally, the parameters are updated using the bias-corrected estimates:

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{2.21}$$

where ϵ is a small constant (e.g., 10^{-8}) to prevent division by zero. Adam dynamically adjusts the learning rate based on the moments, making it more robust and efficient for large datasets and noisy gradients.

There are several types of neural networks designed for different kinds of data. Fully connected neural networks, also known as feedforward neural networks, connect every neuron in one layer to every neuron in the next layer and work well with structured data. However, they may not be effective for capturing spatial or temporal relationships in data, which are crucial for tasks involving images or time series.

Convolutional Neural Networks (CNNs) are specialized neural networks primarily used for image data [98]. CNNs leverage the spatial structure of images, making them particularly effective for tasks like image classification and object detection. The architecture of a CNN typically consists of convolutional layers, pooling layers, and fully connected layers.

In the convolutional layer, the network applies a set of filters (or kernels) to the input image. Each filter convolves over the image, producing a feature map that highlights specific features such as edges, textures, or patterns. The convolution operation can be mathematically expressed as:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \tag{2.22}$$

where $S(i,j)$ is the output feature map, $I(m,n)$is the input image, and $K(i-m, j-n)$ is the filter kernel.

To introduce non-linearity into the model, an activation function (e.g., ReLU) is applied to the feature map:

$$A(i,j) = f(S(i,j)) = \max(0, S(i,j)) \tag{2.23}$$

Following the convolutional layers, pooling layers are employed to reduce the dimensionality of the feature maps and retain the most important information. Max pooling is a common technique, defined as:

$$P(i,j) = \max_{m,n \in R} A(m,n) \tag{2.24}$$

where R is the region over which the max operation is applied, typically a 2×2 grid. This process reduces the size of the feature maps, helping to mitigate overfitting and computational complexity.

After several convolutional and pooling layers, the output is flattened into a one-dimensional vector and passed to fully connected layers for final classification. In these layers, the same principles of forward propagation and weight updates apply as in standard feedforward networks.

The output layer of a CNN often employs a softmax activation function for multi-class classification tasks, producing a probability distribution over the classes:

$$\hat{y}_k = \frac{e^{z_k}}{\sum_j e^{z_j}} \tag{2.25}$$

where z_k is the raw score (logit) for class k, and $\hat{y}_k$ is the predicted probability for that class.

Neural networks, particularly CNNs, have achieved state-of-the-art performance in a wide range of applications, including medical image analysis, where they can assist in tasks such as tumor detection and disease classification. The flexibility and scalability of neural networks make them powerful tools for both research and practical applications in fields ranging from healthcare to autonomous systems.

2.2.5. Fourier Transformation and Mel-Frequency Cepstral Coefficients (MFCCs)

Fourier transformation is a mathematical technique that transforms a time-domain signal into its frequency-domain representation [99, 100, 101]. Mathematically, the Fourier transform $X(f)$ of a continuous signal $x(t)$ is defined as:

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt \tag{2.26}$$

This transformation decomposes the signal $x(t)$ into its constituent sinusoidal components, allowing us to analyze the frequency content of the signal. The output of the Fourier transform is typically represented as a complex spectrum, where the magnitude $|X(f)|$ indicates the amplitude of each frequency component, and the phase $\angle X(f)$ provides information about the timing of those components [100, 101]. In the context of audio signals, the Fourier transform helps identify the dominant frequencies present in a sound, which is crucial for tasks such as speech recognition, music analysis, and pathological voice analysis [102, 103, 104, 105].

Mel-Frequency Cepstral Coefficients (MFCCs) are a widely used feature extraction technique in audio and speech processing. They are derived from the power spectrum of a sound signal and are designed to mimic the human ear's perception of sound. The human auditory system is more sensitive to changes in frequency at lower frequencies than at higher frequencies, which the Mel scale accounts for [102, 103, 104, 105].

To obtain MFCCs, the audio signal first undergoes a pre-emphasis process, where it is passed through a high-pass filter to amplify higher frequencies and improve clarity. Next, the signal is divided into overlapping frames to capture short-term variations, and each frame is multiplied by a window function, such as a Hamming or Hanning window, to minimize discontinuities at the edges. The Fourier transform is then applied to each frame to obtain the frequency spectrum $X(f)$ [100, 101].

The frequency spectrum is processed through a Mel filter bank, which consists of overlapping triangular filters designed according to the Mel scale. The Mel frequency f_m is defined by the following relationship, which converts linear frequencies f to the Mel scale:

$$f_m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.27}$$

After passing through the Mel filter bank, the logarithm of the filtered values is computed to compress the dynamic range, resulting in the log-mel spectrum $L(f_m)$:

$$L(f_m) = \log\left(Mel(f_m)\right) \tag{2.28}$$

Finally, the Discrete Cosine Transform (DCT) is applied to the log-mel spectrum to obtain the MFCCs C_n, which serve as a compact representation of the spectral features of the audio signal:

$$C_n = \sum_{m=1}^{M} L(f_m) \cdot \cos\left(\frac{\pi n}{M}(m - \frac{1}{2})\right), \quad n = 1, 2, \ldots, N \tag{2.29}$$

MFCCs are particularly effective in capturing the timbral characteristics of audio signals, making them valuable for various applications, including speech recognition, music genre classification, and the analysis of pathological voices. By incorporating MFCCs as features in machine learning models, it is possible to improve the accuracy of classification tasks related to audio data [102, 103, 104, 105].

2.2.6. Augmentation Techniques

In the realm of imbalanced datasets, upsampling techniques are essential for enhancing the performance of machine learning models. Among the most used methods are SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) [106, 107]. Both techniques aim to mitigate class imbalance by generating synthetic samples for the minority class, thereby enriching the training dataset and improving the model's ability to generalize.

SMOTE operates on the principle of creating synthetic instances by interpolating between existing minority class samples. To generate new synthetic samples for a specific minority instance x_i, SMOTE first identifies its k nearest neighbors from the same class, denoted as $N_k(x_i)$. For each neighbor $x_j \in N_k(x_i)$, a new synthetic instance x_{new} is produced using the formula:

$$x_{new} = x_i + \lambda \cdot (x_j - x_i) \tag{2.30}$$

Here, λ is a random value sampled from the interval $[0, 1]$. This interpolation allows SMOTE to create new instances that lie on the line segment connecting x_i and its neighbors, effectively increasing the diversity of the minority class without simply replicating existing samples [106].

In contrast, ADASYN enhances the SMOTE approach by adaptively determining the number of synthetic samples to generate based on the complexity of learning for each minority instance. This adaptability is crucial for focusing on instances that are more challenging for the classifier to learn. For each minority instance x_i, ADASYN first calculates the density D_i of its nearest neighbors, which reflects the proximity of majority class instances. This density is given by:

$$D_i = \frac{1}{k} \sum_{j=1}^{k} I(y_j \neq y_i) \tag{2.31}$$

where I is an indicator function that returns 1 if the class label y_j differs from y_i, indicating the presence of a majority class instance. Subsequently, the number of synthetic samples n_i to be generated for each minority instance x_i is computed based on its density relative to the total density of all minority instances:

$$n_i = \frac{D_{\text{total}} - D_i}{D_{\text{total}}} \cdot N_{\text{desired}} \tag{2.32}$$

In this equation, D_{total} represents the cumulative density of all minority instances, and N_{desired} is the total number of synthetic samples intended for the dataset [107]. By employing these techniques, SMOTE and ADASYN not only bolster the representation of the minority class but also enhance the learning capabilities of classifiers by providing a more varied and informative training dataset. The adaptability of ADASYN allows it to concentrate on generating samples in regions where the minority class is underrepresented, significantly improving the model's ability to recognize and classify minority instances effectively [106, 107].

In clinical datasets, Gaussian Noise Upsampling (GNUS) is another viable augmentation technique, particularly when dealing with small, imbalanced datasets, which are common in clinical settings. Unlike SMOTE and ADASYN, GNUS involves adding Gaussian noise to existing minority class instances to create synthetic samples [108]. This technique works by slightly perturbing the original instances to enhance the diversity of the minority class without generating entirely new samples.

Clinical datasets present unique challenges in machine learning, including high class imbalance, a small number of samples, and a large number of features, often with missing values. While imputation techniques and feature selection can address the latter issues, augmentation techniques like GNUS help in handling the imbalance problem. GNUS has

shown promising results in some cases, outperforming SMOTE and ADASYN, particularly when clinical datasets are small and difficult to augment effectively with more complex techniques [108].

Moreover, studies have revealed that augmentation may not always improve classification performance, and in some cases, models perform equally well without augmentation [108]. This highlights the fact that augmentation techniques must be carefully selected based on the nature of the dataset and the specific machine learning task. Nevertheless, GNUS generally enhances classification performance by introducing variability into the minority class, making it a strong contender for clinical datasets where other augmentation techniques might not be as effective.

2.2.7. Conventional, Batch, Layer and Adaptive Normalization

Normalization techniques are essential in neural network training, helping improve convergence and stability by ensuring that activations across layers maintain stable distributions during the training process. This subsection covers four important normalization methods: Conventional Normalization (Z-Score), Batch Normalization, Layer Normalization, and Adaptive Normalization [109, 110, 111, 112, 113].

Normalization techniques are essential in neural network training, helping improve convergence and stability by ensuring that activations across layers maintain stable distributions during the training process. Before diving into the specifics of normalization techniques like Batch, Layer, and Adaptive Normalization, it's important to understand a more conventional approach, standardization.

Standardization or Z-score normalization (conventional normalization) is the process of rescaling the features of a dataset so that they have a mean of zero and a variance of one [109, 110]. For a given feature vector $x_1, x_2, \ldots, x_n$, the standardized version $\hat{x}_i$ is computed as:

$$\hat{x}_i = \frac{x_i - \mu}{\sigma} \tag{2.33}$$

where μ is the mean and σ is the standard deviation of the feature values. Standardization ensures that each feature contributes equally to the model during training, avoiding situations where features with larger values dominate the learning process.

While standardization is applied globally to features before training, more advanced normalization methods like Batch, Layer, and Adaptive Normalization operate within the layers of neural networks, addressing issues such as internal covariate shift and enabling faster and more stable training.

Batch Normalization is one of the most widely used techniques to mitigate internal covariate shift, which occurs when the distribution of network activations shifts during training [111]. Batch Normalization works by normalizing the activations of a given layer across a mini-batch of data, such that the activations have zero mean and unit variance. Specifically, for a mini-batch of activations $x_1, x_2, \ldots, x_m$, the mean μ_B and variance σ_B^2 are computed as follows:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad \sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2 \tag{2.34}$$

These activations are then normalized:

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{2.35}$$

where ϵ is a small constant added for numerical stability. After normalization, a learnable scale parameter γ and a shift parameter β are applied:

$$y_i = \gamma \hat{x}_i + \beta \tag{2.36}$$

Batch Normalization not only speeds up training by stabilizing the distribution of activations but also allows for higher learning rates and acts as a regularizer, reducing the need for dropout [111].

Layer Normalization, on the other hand, normalizes the activations within each individual data sample instead of across a mini-batch [112]. This makes it particularly useful for small batch sizes or in recurrent neural networks (RNNs). The mean μ_L and variance σ_L^2 are calculated across the neurons in a single layer for each data point:

$$\mu_L = \frac{1}{H}\sum_{i=1}^{H} x_i, \quad \sigma_L^2 = \frac{1}{H}\sum_{i=1}^{H}(x_i - \mu_L)^2 \tag{2.37}$$

The activations are normalized according to:

$$\hat{x}_i = \frac{x_i - \mu_L}{\sqrt{\sigma_L^2 + \epsilon}} \tag{2.38}$$

Similar to Batch Normalization, learnable parameters γ and β are applied after normalization:

$$y_i = \gamma \hat{x}_i + \beta \tag{2.39}$$

Layer Normalization is particularly effective for handling sequence data in RNNs or transformers, where maintaining a consistent distribution of activations across sequences is crucial for model performance [112].

Adaptive Normalization, by Ruhland et al., is derived from Batch Normalization, incorporating additional flexibility by using learnable parameters for both the mean and variance [113]. This approach allows for a dynamic adjustment of these statistics during training, making the model more adaptable to different datasets or layers.

In Adaptive Normalization, the mean $M_{j'k'l'}$ and variance $V_{j'k'l'}$ are calculated using two sets of learnable weights, $w^0_{j'k'l'}$ and $w^1_{j'k'l'}$, rather than directly computing them from the data:

$$M_{j'k'l'} = w^0_{j'k'l'} \cdot w^1_{j'k'l'} \tag{2.40}$$

$$V_{j'k'l'} = \max\left\{0, w^0_{j'k'l'} \cdot w^1_{j'k'l'}\right\} \tag{2.41}$$

The rectified linear function (ReLU) ensures that the variance remains non-negative. The activations are normalized as:

$$Y_{ijkl} = \frac{X_{ijkl} - M_{j'k'l'}}{\sqrt{V_{j'k'l'} + \epsilon}} \cdot \gamma_{j'k'l'} + \beta_{j'k'l'} \tag{2.42}$$

The weight $w^0_{j'k'l'}$ is initialized to the mean of the dataset and $w^1_{j'k'l'}$ is initialized to 1, respectively. This dynamic approach enables more flexible and precise control over the normalization process, allowing the model to adapt during training without being constrained by fixed statistics. Additionally, Adaptive Normalization can handle a variety of data types, including one-dimensional, two-dimensional (images), and even higher-dimensional data, by adjusting the indices j, k, and l accordingly. To fasten the training the first normalization layer is initialized by the mean and variance of the respective dataset [113].

While Batch Normalization normalizes activations across a mini-batch and Layer Normalization normalizes each data point individually, Adaptive Normalization strikes a balance by introducing additional learnable parameters, making it highly versatile and suitable for various neural network architectures.

In summary, Batch Normalization, Layer Normalization, and Adaptive Normalization each serve different purposes in stabilizing neural networks during training, with Adaptive

Normalization offering an additional degree of flexibility that can be crucial in more complex models.

2.3. Model Evaluation, Calibration and Causal Discovery

This section discusses key concepts related to the evaluation and calibration of machine learning models, which are crucial for ensuring reliable predictions in healthcare applications.

2.3.1. Evaluation Metrics

To evaluate the performance of both traditional machine learning models and neural networks, a set of commonly used classification metrics were employed. A good starting point for understanding these metrics is the confusion matrix, which provides a comprehensive overview of the model's predictions [114].

The confusion matrix is a table that summarizes the performance of a classification algorithm by displaying the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). It can be represented as follows:

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

From this matrix, several important metrics can be derived [114]. Accuracy is calculated as the proportion of correct predictions over all predictions made and is given by the formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.43}$$

While accuracy is a widely used metric, it can be misleading, especially in cases with imbalanced class distributions [115, 116]. For example, in a dataset where only 10% of the samples belong to the positive class, a model predicting all samples as negative would achieve an accuracy of 90%, despite failing to identify any true positives.

To gain deeper insights into the model's performance, Sensitivity, also known as Recall or the True Positive Rate, is defined as:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2.44}$$

This metric indicates how well the model identifies actual positive cases, minimizing false negatives [115, 116]. Conversely, Specificity, or the True Negative Rate, is defined as:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2.45}$$

Specificity reflects how well the model identifies negative cases, thereby minimizing false positives [115, 116]. Both sensitivity and specificity provide essential insights into the model's behavior, allowing for a more nuanced understanding of its performance [115, 116].

Precision, also referred to as the Positive Predictive Value, measures the proportion of correctly predicted positive observations to the total predicted positives [115, 116]. It can be calculated using the formula:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.46}$$

Precision becomes particularly important when the costs of false positives are high [115, 116]. To combine the strengths of both precision and sensitivity into a single metric, the F1 Score is used, which is defined as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{2.47}$$

The F1 Score provides a balanced view of the model's performance, especially in scenarios where there is an uneven class distribution [115, 116].

Another critical metric is the ROC-AUC (Receiver Operating Characteristic - Area Under the Curve), which measures the area under the ROC curve [117]. The ROC curve plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) at various threshold settings [117]. The ROC-AUC is calculated as:

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(FPR)\, d(\text{FPR}) \tag{2.48}$$

A higher ROC-AUC value indicates a better balance between sensitivity and specificity [117]. However, it is important to note that while ROC-AUC provides valuable insights, it may not be the most reliable metric in cases of imbalanced datasets [117].

Given these considerations, while ROC-AUC is a significant metric, precision and recall are crucial for evaluating predictive performance, especially in medical datasets characterized by class imbalances [115, 116, 117]. Thus, this study emphasizes the importance of employing a combination of metrics to ensure a comprehensive evaluation that accurately reflects the model's performance in light of the data's complexities.

2.3.2. Optimal Cutoff Analysis

The selection of the cutoff value x_{cutoff} was defined using the following formula:

$$x_{cutoff} = \max\left(1 - r_{fp} + r_{tp}\right) \tag{2.49}$$

Where:

- r_{fp} represents the rate of false positives,
- r_{tp} represents the rate of true positives.

This formula is a heuristic method to balance the trade-off between sensitivity and specificity. By maximizing $1 - r_{fp} + r_{tp}$, we aim to minimize the impact of false positives while maximizing the detection of true positives. This approach provides a systematic way to identify the threshold that achieves the optimal classification performance at a given state of the model.

It is important to note that while this heuristic approach helps identify a cutoff that balances various performance metrics, it does not guarantee better performance in all cases. The choice of cutoff is highly dependent on the specific dataset and model behavior, and there may be scenarios where a different threshold selection strategy yields better results. As such, the selected cutoff should be interpreted as a practical estimate rather than an absolute optimal solution.

2.3.3. DeLong Statistics

To evaluate and compare the performance of the best neural networks trained on the complete and non-invasive feature sets, the DeLong statistic was employed [118]. This statistical method allows for the comparison of two correlated receiver operating characteristic (ROC) curves, which is essential when assessing models built on overlapping data or subsets of features.

The DeLong test specifically evaluates whether the area under the ROC curve (AUC) for one model is significantly different from that of another model. By utilizing this method, we were able to quantify whether the performance difference between the model trained on the full feature set and the one trained only on non-invasive features was statistically significant.

This analysis provided a robust metric for determining the viability of using non-invasive features alone for predictive modeling. If the DeLong test indicated no significant difference between the two models, it would suggest that a model trained solely on non-invasive features could potentially replace the one trained on the full feature set, thereby reducing the need for more invasive data collection methods.

2.3.4. Pivotframe Visualization

A Pivotframe visualization is a tool used to display relationships between different features in a dataset by organizing them into a grid format. It uses a heatmap, where colors represent the strength or significance of the relationships between feature combinations and an outcome of interest. Each cell in the grid corresponds to a unique combination of feature values, and the color of the cell indicates the strength of the relationship, such as the frequency of an outcome or a model's prediction.

One significant use of Pivotframes is in validating neural networks. When trained neural networks are tested using simulated data, Pivotframes can help assess whether the model has learned something meaningful or merely captured biases from the dataset. By visualizing the predictions of the neural network over various feature combinations, a Pivotframe helps identify whether the network generalizes well or if it is overfitting or underfitting based on the dataset's inherent biases.

For example, if a neural network's predictions consistently reflect expected patterns in the data across different feature combinations in the Pivotframe, this indicates meaningful learning. On the other hand, if the model's predictions seem random or overly simplistic (e.g., consistently predicting a dominant class regardless of feature variations), it may suggest that the model has not truly captured the complexities of the data and has instead learned a bias. Thus, Pivotframe visualizations serve as a critical diagnostic tool for evaluating the robustness and generalization of machine learning models.

2.3.5. Probability Calibration Framework GUESS

In clinical decision-making, machine learning models are increasingly applied to predict outcomes in areas such as cancer survival and drug resistance. While these models can be highly accurate, their output is often not easily interpretable as class probability estimates. This is particularly concerning in clinical settings, where understanding probabilistic outcomes is critical for making informed decisions [119].

To address this limitation, Schwarz et al. and Heider et al. introduced a novel method called GUESS, designed to project machine learning scores into well-calibrated probability estimates [119]. In clinical applications, it is not enough to optimize models solely based on discrimination (i.e., how well a model distinguishes between classes); they must also provide accurate calibration—the ability to project scores that reflect true probability estimates.

The authors compared GUESS with two state-of-the-art calibration techniques: Histogram Binning and Bayesian Binning in Quantiles. Their findings demonstrated that GUESS achieved calibration performance comparable to these methods while maintaining high accuracy in class discrimination. Furthermore, GUESS showed superior performance in

small datasets, making it a suitable option for various clinical datasets, which often suffer from limited data availability.

By enabling more interpretable and reliable probability estimates, GUESS contributes to the acceptance and adoption of machine learning models in clinical decision support systems, especially in sensitive applications like disease prediction and risk assessment.

2.3.6. Causal Framework

Causal discovery involves determining the underlying cause-and-effect relationships between variables in a dataset. One algorithm in this domain is the Fast Causal Inference (FCI) algorithm, which infers a Partial Ancestral Graph (PAG) representing a Markov representation of all potential causal graphs. The FCI algorithm achieves this by applying conditional independence tests to the data, which allows it to handle scenarios with latent variables and selection bias [120]. The FCI algorithm is based on constraint-based methods and relies heavily on conditional independence tests. These tests are used to iteratively refine the possible causal structures by identifying where conditional independencies occur.

The Fisher's Z test assesses the partial correlation between two variables given a set of conditioning variables [121, 122]. The test statistic is computed as follows:

$$Z = \sqrt{n - |C| - 3} \times 0.5 \times \ln\left(\frac{1 + \hat{\rho}_{XY|C}}{1 - \hat{\rho}_{XY|C}}\right) \tag{2.50}$$

Where n is the sample size, $|C|$ is the number of conditioning variables, and $\hat{\rho}_{XY|C}$ is the partial correlation coefficient between variables X and Y, conditioned on C. A high absolute value of Z indicates that X and Y are likely dependent given C.

The chi-squared conditional independence test, another popular method, is typically used for categorical data [123]. It assesses the independence between two variables by comparing observed and expected frequencies. The test statistic is given by:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{2.51}$$

Where O_i are the observed frequencies and E_i are the expected frequencies under the null hypothesis of independence. This test evaluates whether the distribution of the data deviates significantly from what is expected if the variables were independent.

Similarly, the G-squared (likelihood ratio) test can be applied to categorical variables [124]. The test statistic is computed as:

$$G^2 = 2 \sum O_i \ln\left(\frac{O_i}{E_i}\right) \tag{2.52}$$

This test provides a measure of how much the observed data diverges from the expected data under the assumption of independence.

For non-parametric CI testing, the kernel-based CI test (KCI) uses kernels to measure dependence between variables [125]. While this method can detect non-linear dependencies, its complexity is cubic in the sample size, which can make it slower for larger datasets. The test statistic is approximated using the Hilbert-Schmidt independence criterion (HSIC):

$$HSIC(X, Y|Z) = \frac{1}{(n-1)^2}\mathrm{Tr}(K_X H K_Y H) \tag{2.53}$$

Where K_X and K_Y are kernel matrices and H is a centering matrix.

The Missing-Value Fisher's Z (MV Fisher-Z) is a variation of the standard Fisher's Z test that deals with datasets containing missing values [121, 122, 126]. It uses techniques such as imputation or listwise deletion to handle the missing data, thereby extending the use of Fisher's Z to more practical real-world datasets.

Each of these conditional independence tests plays a crucial role in identifying causal relationships within the data, allowing the FCI algorithm to robustly infer causal structures even when latent variables are present or when data is missing. Understanding the advantages and limitations of each CI test is essential for selecting the most appropriate method for a given dataset.

The output of the FCI algorithm is typically a Partial Ancestral Graph (PAG), which represents the possible causal relationships in a dataset while taking into account latent confounders and selection bias [127]. Unlike traditional directed acyclic graphs (DAGs), PAGs are more flexible, allowing for the presence of hidden variables that cannot be directly observed or measured [128].

PAGs utilize a combination of edge types to encode causal relationships. Directed edges, represented as $X \to Y$, indicate that X is a direct cause of Y, assuming there are no latent confounders. In contrast, bidirected edges, denoted as $X \leftrightarrow Y$, suggest the presence of a latent common cause, meaning both X and Y share a hidden confounder. Additionally, circle-marked edges, expressed as $X \circ\!\!-\!\!\circ Y$, represent uncertainty regarding the directionality of causality, implying that further data or tests are needed to determine the exact causal direction. Furthermore, edges with one side as an arrow and the other as a circle, illustrated as $X \circ\!\!\to Y$, indicate that while X may influence Y, there is uncertainty about whether there is a direct causal relationship due to potential hidden variables.

PAGs are particularly valuable in real-world scenarios where data may be incomplete or where hidden confounders could influence the relationships between variables. The different edge markings in a PAG indicate that the underlying causal relationships are not fully determined, making these graphs crucial tools in exploratory causal analysis. By carefully

interpreting the edges in a PAG, researchers can formulate hypotheses about the causal structure of the data and design further studies to refine their understanding.

3. Material and Methods

To achieve accurate disease prediction and meaningful causal insights, various datasets, tools, and machine learning models were employed throughout this study. This section provides an overview of the modules, data sources, model architectures, and analytical methods utilized. The table A.1 lists the key modules used for the implementation of the machine learning models and analysis in this thesis, along with their respective descriptions and versions.

3.1. Datasets Overview

The UCI Machine Learning Repository and Kaggle databases were scanned for suitable datasets. A total of 20 datasets were evaluated and 8 were found to be adequate for the task.

The UCI Machine Learning Repository offers a wide range of datasets from various scientific domains, frequently used for machine learning purposes. It includes datasets relevant to disease prediction and medical data analysis.

Kaggle is a platform providing numerous datasets commonly used in data science competitions and machine learning projects. The datasets include extensive clinical and medical information, suitable for non-invasive disease prediction.

3.1.1. Heart Disease Dataset

The Heart Disease dataset used in this study is the *Heart Failure Prediction Dataset*, available on Kaggle [129]. It was created by combining five independent datasets: Cleveland, Hungarian, Switzerland, Long Beach, and Statlog. These datasets were originally collected in various medical facilities in the United States and Europe, with clinical and demographic data obtained through patient records and examinations. The merged dataset contains 918 samples and 12 attributes, making it a comprehensive heart disease dataset for research.

The key features include **Age**, representing the patient's age in years, with 67.2% of the individuals being between 40 and 60 years old; **Sex** indicating the patient's gender (M for male and F for female), with males comprising 78.9% of the sample population; and **ChestPainType**, classified into Typical Angina (TA), Atypical Angina (ATA),

Non-Anginal Pain (NAP), and Asymptomatic (ASY), reflecting different forms of chest pain. Additional features such as **RestingBP**, **Cholesterol**, and **MaxHR** (measured with photoplethysmography) identify cardiovascular health, while **ExerciseAngina** and **ST_Slope** capture the patient's response to physical exertion [130]. The target variable, **HeartDisease**, indicates the presence or absence of heart disease. This dataset thus provides a broad view of the demographic, lifestyle, and clinical factors associated with heart disease [129].

3.1.2. Breast Cancer Dataset

The Breast Cancer dataset, titled *Breast Cancer Prediction*, is publicly available on Kaggle [131]. The data was collected from the University of Calabar Teaching Hospital over a period of 24 months, from January 2019 to August 2021, and consists of 213 observations. This dataset was curated to investigate the relationship between various diagnostic attributes of patients and their breast cancer diagnosis [131].

The dataset includes the following features: **Age**, representing the patient's age at the time of diagnosis; **Menopause**, indicating whether the patient is pre- or postmenopausal, with 0 meaning postmenopausal and 1 meaning premenopausal; and **Tumor size**, measuring the excised tumor size in centimeters. Other attributes, such as **Involved nodes** and **Metastatic**, describe the presence of cancer in axillary lymph nodes and its spread to other parts of the body, respectively.

Additionally, the **Breast** feature specifies whether the cancer affects the left or right breast, while **Breast quadrant** identifies the location of the tumor within one of the four sections of the breast. The **History** attribute captures any personal or family history of cancer. The **Diagnosis result** serves as the target variable, where 0 indicates a benign diagnosis and 1 indicates a malignant diagnosis. Overall, this dataset offers a detailed view of clinical factors relevant to breast cancer diagnosis and patient history.

3.1.3. Alzheimer Disease Dataset

The Alzheimer Disease dataset, named *Alzheimer Disease Dataset*, is accessible on Kaggle and includes data for 2,149 patients [132]. The data was collected over several years, comprising demographic details, lifestyle factors, medical history, clinical measurements, and cognitive assessments to study the risk and progression of Alzheimer's Disease [132].

Key demographic features include **Age**, which ranges from 60 to 90 years, and **Gender**, coded as 0 for males and 1 for females. **Ethnicity** and **EducationLevel** categorize patients based on ethnicity (e.g., Caucasian, African American) and education levels (e.g., High School, Bachelor's). The dataset also captures lifestyle factors such as **BMI**, **Smoking**, **AlcoholConsumption**, and weekly **PhysicalActivity**.

Medical history attributes like **FamilyHistoryAlzheimers**, **CardiovascularDisease**, **Diabetes**, and **Hypertension** indicate comorbidities and risk factors. Clinical measurements, including **SystolicBP**, **CholesterolTotal**, and **CholesterolHDL**, provide insight into cardiovascular health.

Cognitive and functional assessments are represented by the **MMSE** (Mini-Mental State Examination) scores and **FunctionalAssessment**, while symptoms such as **Forgetfulness**, **Confusion**, and **Disorientation** capture cognitive decline. The target variable **Diagnosis** indicates the presence or absence of Alzheimer Disease. This dataset enables comprehensive analysis of the factors associated with Alzheimer's progression.

3.1.4. Depression Dataset

The *Indicators of Anxiety or Depression Based on Reported Frequency of Symptoms During Last 7 Days* dataset, sourced from Kaggle, contains information collected through the Household Pulse Survey conducted by the U.S. Census Bureau [133]. The survey aims to evaluate the social and economic impacts of COVID-19 on American households, focusing on mental health metrics. Data is gathered through weekly internet-based questionnaires, making it one of the most current datasets available for analyzing trends in anxiety and depression during the pandemic.

The dataset provides weighted estimates that reflect the population's demographic structure, segmented by variables such as **age**, **sex**, **race/ethnicity**, **education**, and **state**. The key features include weekly prevalence of mental health indicators like **Symptoms of Anxiety** or **Depression**, represented as percentages. Each record includes the confidence interval bounds and a unique identifier for the specific survey **Phase** and **Time Period**, ensuring accurate tracking of changes over time.

This dataset facilitates the examination of mental health trends across various demographic groups, highlighting the impact of external factors on anxiety and depression.

3.1.5. Chronic Kidney Disease Dataset

The Chronic Kidney Disease dataset is sourced from the UCI Machine Learning Repository and was collected over two months in India [134]. It aims to support healthcare professionals in predicting chronic kidney disease [134].

This dataset includes 24 features, encompassing both numerical and categorical variables. Key numerical features include age (in years), blood pressure (mm/Hg), blood urea, serum creatinine, sodium, potassium, and hemoglobin levels. Categorical features include specific gravity, albumin, red blood cells, pus cells, presence of pus cell clumps and bacteria, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, and anemia.

The target variable, *class*, indicates whether a patient has chronic kidney disease or not. The dataset features a balanced age distribution, with ages ranging from young adults to the elderly, and a roughly equal gender representation. This diversity enhances the robustness of predictive models developed from the data.

3.1.6. Multiple Sclerosis Disease Dataset

The Multiple Sclerosis Disease dataset, sourced from Kaggle, originates from a cohort study conducted at the National Institute of Neurology and Neurosurgery in Mexico City (2006-2010) [135]. It focuses on Mexican mestizo patients diagnosed with Clinically Isolated Syndrome (CIS) and is vital for analyzing multiple sclerosis (MS), the most common demyelinating disease affecting the central nervous system [135].

This prospective cohort study involved patients newly diagnosed with CIS. MS is characterized by the damage to the insulating covers of nerve cells, resulting in symptoms such as visual disturbances, muscle weakness, and coordination issues. While the exact cause remains unclear, factors like genetics and environmental influences are believed to contribute to its development.

The dataset includes various features: patient ID, age, schooling, gender, breastfeeding history, varicella status, initial symptoms, and MRI results. The target variable classifies patients into groups—CDMS and non-CDMS—providing insights into conversion predictors from CIS to MS.

The dataset is publicly accessible under the CC BY 4.0 license and contains critical information for predicting disease progression and improving patient management strategies.

3.1.7. Thyroid Cancer Recurrence Dataset

The Differentiated Thyroid Cancer Recurrence dataset, sourced from the UCI ML Repository, comprises 13 clinicopathologic features used to predict the recurrence of well-differentiated thyroid cancer [136]. Over a 15-year period, data were collected from 383 patients, each followed for a minimum of 10 years [136].

This dataset is essential for understanding the factors influencing recurrence in thyroid cancer and can aid in developing predictive models for patient management.

3.1.8. Pathological Voice Dataset

The VOICED dataset, sourced from Kaggle, comprises 208 voice samples, including 150 pathological and 58 healthy voices [137]. These samples were clinically verified using the SIFEL protocol. The dataset encompasses 13 clinicopathologic features aimed at

understanding voice disorders. The collected data includes demographic information such as age and gender, as well as lifestyle habits, including smoking status and alcohol consumption.

Participants were evaluated using two specific medical questionnaires: the Voice Handicap Index (VHI), which assesses the impact of voice disorders on quality of life, and the Reflux Symptom Index (RSI), which evaluates the severity of reflux symptoms. The voice recordings were made under controlled conditions at a sample rate of 8000 Hz, focusing on the vowel 'a'.

The study involved adults aged 18 to 70, excluding individuals with certain health conditions like respiratory infections or neurological disorders. The data collection occurred over a period of one year, from May 16, 2016, to May 15, 2017. The recordings were performed in a quiet environment, ensuring minimal background noise and appropriate humidity levels.

This dataset serves as a valuable resource for researchers and clinicians working on voice disorders, providing insights into the characteristics and implications of both healthy and pathological voices [137].

3.1.9. Diabetes Dataset

This study utilizes data from the Heinz-Nixdorf-Recall (HNR) cohort, a population-based sample comprising 4,814 participants aged 45–75 years (2,419 female) [10, 138, 139]. Key features in the dataset include demographic and lifestyle variables such as age, sex, height, weight, waist circumference, alcohol consumption (beer), and smoking status. Men in the cohort displayed significantly higher average weight and waist circumference than women [10]. The overall mean BMI of 27.9 kg/m^2 suggests that the cohort predominantly consists of overweight or obese individuals, a critical factor for diabetes risk analysis [10].

3.2. Data Analysis

This section provides an overview of the data preprocessing, machine learning models tested, model architectures, and evaluation strategies applied to the selected datasets. A variety of preprocessing techniques were employed to ensure data quality and consistency before implementing a diverse set of machine learning models.

3.2.1. Data Preprocessing and Augmentation

Data preprocessing was conducted to prepare the datasets for machine learning. Categorical variables were encoded into numerical form using label encoding. Numerical variables were standardized using z-score normalization, which was applied only after addressing missing values and outliers [109, 110].

For handling missing values, mean imputation was used for numerical variables and mode imputation for categorical variables, depending on the nature and proportion of missing entries [140]. Features with excessive missing values were removed to maintain data integrity. Additionally, outliers were identified and handled using the Interquartile Range (IQR) method to prevent them from skewing the results [141].

A consistent train-validation-test split was applied to each dataset, where 80% of the data was used for training, and the remaining 20% was equally divided into validation and testing sets. Stratified sampling was employed to maintain the class distribution balance across the splits, ensuring proportional representation of minority and majority classes [142].

In cases where a model exhibited signs of overfitting on the majority class, techniques such as SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic Sampling), or adding Gaussian noise to upsample (GNUS) the minority class were employed to address class imbalance [106, 107, 108].

The pathological dataset includes audio samples for each patient, which underwent preprocessing. These audio samples were transformed into spectrograms to train convolutional neural networks. Additionally, a Fourier transform was applied to the audio samples to obtain the Mel-frequency cepstral coefficients, which represent the short-term power spectrum of sound [99]. These coefficients were then incorporated as features into the dataset.

3.2.2. Model Training and Evaluation

After preprocessing, a pipeline of traditional machine learning models was constructed and evaluated. The models tested included:

- Support Vector Machine (SVM)
- k-Nearest Neighbors (KNN)
- Random Forest
- XGBoost
- Gradient Boosting Classifier

Each model underwent hyperparameter tuning using randomized parameter search from SKlearn with predefined parameter grids [143]. The pipeline utilized 10-fold cross-validation to ensure robust performance and generalization. The best hyperparameters for each model were selected based on average performance metrics, such as accuracy, sensitivity, specificity, and F1-score, across the validation folds.

For each dataset, the pipeline was executed twice: once using the full set of features and once using only the non-invasive features. This approach enabled the comparison of model performance when restricting the input to non-invasive measurements alone.

To further enhance model performance, conventional normalization was applied to the training data of every model in the pipeline using the scikit-learn library [143]. It is crucial to calculate the mean and variance of the respective training set. This approach prevents information leakage from the validation and test sets, which would otherwise skew the evaluation of the model's performance. After fitting on the training data, the same mean and variance are used to transform both the validation and test sets, ensuring that these datasets are scaled in a consistent manner. This practice helps maintain the integrity of the model evaluation process and provides a fair assessment of the model's generalization capabilities.

3.2.3. Feature Importance Analysis

In addition to model evaluation, a feature importance analysis was conducted using the Random Forest model from the pipeline.

By examining the importance scores of each feature, the analysis aimed to compare the predictive significance of invasive versus non-invasive features.

The feature importance analysis revealed which variables had the greatest impact on the models predictions, providing valuable insights into the relative importance of different feature types. This comparison enabled the identification of non-invasive features that might be used effectively in disease prediction, reducing the need for invasive testing while still maintaining accurate predictions.

3.2.4. Neural Network Architecture

For each dataset, two separate neural network models were trained, as with the traditional model pipeline explained previously:

1. A model trained on the full feature set, which included both non-invasive and invasive features.
2. A model trained on a reduced feature set consisting solely of non-invasive features.

This dual-model approach was employed to evaluate whether it is feasible to reduce the feature set to only non-invasive features without significantly compromising performance. If the model trained on non-invasive features performs comparably to the one trained on the complete feature set, this would indicate the potential for building predictive models based only on non-invasive measurements, thus avoiding the need for more intrusive tests.

Different neural network architectures were tested in this process. These included varying the number of fully connected layers, experimenting with different activation functions such as ReLU and Tanh, as well as testing different output layer configurations, including

Sigmoid and Softmax functions. Additionally, various hidden layer sizes were explored to assess their impact on model performance.

The architecture and optimization strategies for these neural networks were handled separately, with a focus on improving model performance through tuning. Notably, an analysis and comparison between different normalization techniques such as batch normalization, layer normalization, and adaptive normalization was conducted to determine their effects on training stability and convergence.

3.2.5. Training, Evaluation, and Hyperparameter Tuning for the Neural Networks

The neural network models were trained using binary cross-entropy loss for binary classification tasks and cross-entropy loss for multiclass classification tasks [144]. Through pytorch, the Adam optimizer was employed to minimize the loss function, and a Reduce on Plateau learning rate scheduler was applied to dynamically adjust the learning rate when the error plateaued [97, 144]. This setup helped stabilize training and prevent the model from stagnating in local minima.

Key hyperparameters, such as the learning rate, number of epochs, batch size, learning rate scheduler factor, learning rate scheduler patience, and hidden layer size, were tuned using optuna [145]. For each dataset, between 10 and 100 optuna trials were conducted depending on the dataset size. A predefined parameter grid was explored during these trials, with validation loss and accuracy used as the metrics to guide the search. The validation set allowed for monitoring of both training and validation loss for each epoch, helping identify any overfitting.

The best trial was selected based on accuracy and AUC scores. After identifying the optimal model configuration, the final model was tested on a held-out test set to evaluate its ability to generalize to unseen data. This ensured that the model did not overfit the training data and maintained strong performance on new, unseen examples. The hyperparameters that produced best performance in the tuning process can be seen in table A.2.

For evaluation on the test set, multiple metrics were computed including roc-auc, sensitivity, specificity, F1 score, and accuracy. Additionally, a confusion matrix and classification report were generated to obtain detailed model performance metrics, enabling comprehensive analysis and interpretation of the models effectiveness [143].

3.2.6. Optimal Cutoff Analysis for All Features and Non-Invasive Models

An optimal cutoff analysis was applied to the classification results obtained from all features in conjunction with the non-invasive models for each dataset. After training and validating

the machine learning models, determining the most appropriate threshold to optimize classification performance was essential.

For each dataset, the cutoff value x_{cutoff} was calculated based on the rates of false positives r_{fp} and true positives r_{tp}. This heuristic method allowed us to balance sensitivity and specificity, which are critical in medical applications where the costs of false negatives and false positives can be significant.

The analysis was conducted on the validation sets from each dataset, ensuring that the optimal cutoff values reflected the models performance under conditions similar to real-world applications. The final cutoff for each non-invasive model was selected based on the evaluation metrics explained before.

3.2.7. Network Comparison

To assess the performance differences between the non-invasive models and the models trained on all features, the DeLong statistic was conducted for each dataset [118]. This statistical method compares two correlated receiver operating characteristic (ROC) curves, allowing for a robust evaluation of the models.

By applying the DeLong test, we aimed to determine whether there was a statistically significant difference in the area under the ROC curve (AUC) between the non-invasive model and the model utilizing the complete feature set. This analysis was crucial for evaluating the viability of relying solely on non-invasive features for predictive modeling.

If the DeLong test indicated no significant difference between the two models, it would imply that the non-invasive model could serve as an effective alternative to the full feature model, potentially simplifying the data collection process and enhancing patient comfort by minimizing the need for invasive procedures. Thus, the results from the DeLong analysis provided valuable insights into the performance of the different modeling approaches across all datasets.

3.2.8. Probabilistic Interpretation of Machine Learning Scores

Probability calibration was performed for both the non-invasive model and the model trained on all features across all datasets by using GUESS [119]. Accurate calibration of the predicted probabilities allows for a more meaningful interpretation of the machine learning scores, which is particularly crucial in clinical settings where decisions are made based on these probabilities.

For each dataset, both models were used to make predictions on the training data. The predicted probabilities from the training set were then used to fit GUESS. Through GUESS, the best likelihood distribution for the calibration was determined. This process involved fitting various probability distributions and identifying the one with the highest likelihood.

To ensure the robustness of the calibration, we visualized the fit functions for the predicted probabilities to check for inconsistencies in the distributions. If any inconsistencies were observed, such as unrealistic drops or fluctuations in the distribution, we selected a different distribution that, while theoretically providing a slightly worse fit, produced more realistic results. This trade-off between theoretical accuracy and practical realism ensured that the calibrated probabilities were suitable for clinical decision-making.

After selecting the final fit function or distribution for each dataset, we calculated the calibration errors (ECE, MCE, LCE). These calibration errors quantify how well the predicted probabilities match the actual outcomes and provide insights into the reliability of the model's predictions.

3.2.9. Pivot Table Analysis for Model Predictions and Calibrated Scores

Pivot tables were created for both models across each dataset to facilitate a detailed comparison of their performance. For each model, two distinct pivot tables were generated: one based on the model's predictions and the other reflecting the calibrated scores obtained through the GUESS framework.

To ensure a robust validation process, the predictions and calibrated scores used for the pivot tables were obtained through a synthetically generated dataset. This synthetic dataset provided an ideal testing ground to validate whether the model had captured any inherent biases from the original dataset or had generalized well. By using this approach, we were able to investigate whether the neural networks were learning meaningful patterns or simply replicating biases present in the training data.

This analysis was particularly important for validating that the neural networks had learned meaningful patterns rather than merely capturing biases present in the dataset. By observing whether the predictions reflected expected relationships across various feature combinations, we could determine if the models had successfully generalized from the training data.

3.2.10. Intervention Analysis

To further explore the predictive and causal capabilities of our models, we conducted an intervention analysis using synthetic data. By keeping one feature constant while randomly uniform-generating others, we aimed to understand how the constant feature influences the models predictions.

For instance, in analyzing heart disease predictions, we held blood pressure constant and incrementally increased its value. This allowed us to observe how these variations impacted the models output probability of heart disease, $P(\text{heartDisease} \mid \text{inputFeatures})$. By

iteratively adjusting individual features, we were able to examine their effects on prediction accuracy and outcomes, which can inform clinical decision-making and intervention strategies.

In our intervention analysis, we compared how altering a single feature impacts predictions: $P(\text{heartDisease} \mid f_0 = x_1, f_1, \ldots, f_N) > P(\text{heartDisease} \mid f_0 = x_2, f_1, \ldots, f_N)$, where f_0 is the constant feature being varied, while others remain fixed. This comparison enabled us to gauge the effect of a feature on the model's output.

To implement this intervention analysis, we employed three distinct approaches:

1. GUESS Calibration: This approach involved calibrated probabilities to calculate the mean calibration for predictions across varying values of the constant feature.
2. Model Scores: Here, we computed mean scores using the raw output probabilities for different values of the constant feature.
3. Threshold Application: In this approach, we applied a threshold to the raw model scores, transforming them into binary outcomes based on whether they exceeded a threshold of 0.5 and calculated the mean.

Through this intervention analysis, we aimed to reveal how specific changes in features could alter predictive outcomes.

3.2.11. Causal Discovery Analysis

To uncover the causal relationships between features in the heart disease and the diabetes dataset, a causal discovery was performed using the Fast Causal Inference (FCI) algorithm [120]. A reduced feature subset was selected to ensure the analysis remained conductible. This step was crucial as causal algorithms, especially those based on conditional independence tests, face performance bottlenecks when dealing with high-dimensional data in combination with a small sample sizes.

The dataset used for this causal analysis was the same training data employed in the neural network models, ensuring consistency across the various methodologies applied in this research. For implementing the FCI algorithm, we utilized the Python-based causallearn library, leveraging its functionality to compute a Partial Ancestral Graph (PAG) from the dataset [126]. The FCI algorithm was executed using default parameters, adhering to standard practices to ensure reproducibility and robustness of the results.

To explore potential variations in the causal graph, the algorithm was run with different conditional independence tests, including Fisher's Z, Chi-squared, G-squared, and kernel-based tests [121, 122, 123, 124, 125]. This allowed us to observe how the choice of test affected the structure of the resulting PAG and provided insights into the sensitivity of the causal discovery process to the underlying statistical assumptions of each test.

4. Results

4.1. Feature Importance Analysis

This analysis identifies the most influential features contributing to disease prediction for each condition, distinguishing between invasive and non-invasive features. It is important to note that this was done exclusively on the training set because the training set was used to train the Random Forest.

The results showed that, for datasets such as Heart Disease, Alzheimer, Depression, Pathological Voice, and Multiple Sclerosis, the majority of the most important features were non-invasive. In contrast, for the Breast Cancer dataset, the invasive features consistently outperformed the non-invasive ones in terms of importance. For Chronic Kidney Disease (CKD) and Differentiated Thyroid Cancer Recurrence, both invasive and non-invasive features played significant roles, with a slight predominance of the invasive features.

Subsequently, the Random Forest classifier was retrained using a reduced set of selected features. The cutoff for feature inclusion was based on their importance scores: any feature with an importance score less than half of the highest importance score was excluded from the model.

Despite the feature selection process, retraining the model did not result in improved performance except for the Pathological Voice Dataset. The model showed a significant improvement after training on the eighteen best features of the analysis. For the other models the predictive accuracy either remained unchanged or deteriorated.

The detailed feature importance rankings for each dataset can be found in the following tables: A.3 (Heart Disease), A.4 (Alzheimer's Disease), A.5 (Breast Cancer), A.6 (Depression), A.7 (Chronic Kidney Disease), A.8 (Multiple Sclerosis), A.9 (Pathological Voice), and A.10 (Differentiated Thyroid Cancer Recurrence).

4.2. Model Performance

The evaluation of the machine learning models revealed that, for every dataset except for Pathological Voice, the neural network achieved the best performance. In the case of the Pathological Voice dataset, however, a Random Forest classifier trained on the top 18 features from the importance analysis outperformed the neural network.

Additionally, the results indicated that, for nearly all datasets, the model trained on non-invasive features outperformed or had similar performance as the model trained on all features.

The ROC-AUC curves for the best models for the respective dataset can be seen in, A.1 (Heart Disease), A.2 (Alzheimer Disease), A.3 (Breast Cancer), A.4 (Chronic Kidney Disease), A.6 (Multiple Sclerosis), A.5 (Differentiated Thyroid Cancer Recurrence).

A complete overview of the performance metrics can be seen in the following tables for their respective dataset. Tables A.11, A.12 for breast cancer. Tables A.13, A.14 for heart disease. Tables A.15, A.16 for CKD. Tables A.17, A.18 for multiple sclerosis. Tables A.19, A.20 for differentiated thyroid cancer recurrence.

4.3. Model Architecture and Augmentation Techniques

In testing various networks on both non-invasive and complete feature sets across all datasets, the most effective architecture consistently consisted of an initial adaptive normalization layer, followed by a sequence of fully connected layers utilizing ReLU activation functions. Adaptive normalization proved to be more effective than layer, batch, and conventional normalization techniques. Additional normalization layers were applied between the fully connected layers to enhance model performance and training speed.

The results of the augmentation techniques on the training set varied among the datasets. For Alzheimer and Chronic Kidney Disease, Gaussian Noise Upsampling demonstrated superior performance. Conversely, for Breast Cancer, SMOTE produced the best results compared to Gaussian Noise Upsampling and ADASYN. In the case of Heart Disease, Gaussian Noise Upsampling again outperformed SMOTE and ADASYN. The analysis for Depression and Multiple Sclerosis did not highlight a need for augmentation techniques since they only led to overfitting for the upsampled class.

The Random Forest model achieved optimal performance on the Pathological Voice dataset with the hyperparameters listed in Table 4.1. Notably, it outperformed the Neural Network, particularly when utilizing the added Mel Frequency Cepstral Coefficients obtained through Fourier transformation. This suggests that the Neural Network may have insufficient samples to learn effectively. Additionally, the Random Forest demonstrated superior results compared to the CNNs trained on various types of spectrograms.

The Random Forest model was trained on the top 18 features identified in the feature importance analysis, leading to a significant improvement in performance, as shown in Table 4.2. Consistent with the other datasets, data augmentation was performed using SMOTE to upsample the minority class in the training set. By focusing on the top 18 features, the model was able to reduce overfitting, resulting in a more balanced performance on the test set.

Tabelle 4.1.: This table enumerates the hyperparameters that were used to train the Random Forest model for the Pathological Voice dataset.

Hyperparameter	**Value**
random_state	42
bootstrap	False
class_weight	None
max_depth	10
max_features	sqrt
min_samples_leaf	2
min_samples_split	15
n_estimators	387

Tabelle 4.2.: This table illustrates the difference in performance metrics for the Random Forest model trained on all features and the model trained on the best 18 features from the importance analysis.

Model	**ROC-AUC**	**Sensitivity**	**Specificity**	**F1-Score**	**Accuracy**
Random Forest	0.677	0.888	0.466	0.813	0.738
Random Forest Top 18	0.859	0.851	0.866	0.884	0.857

4.4. Cutoff Analysis

As mentioned in the previous chapter, the cutoff analysis is a heuristic approach that does not guarantee improved results. For all datasets analyzed, there were no recorded improvements in the evaluation metrics; they either remained the same or were slightly worse compared to the default threshold of 0.5. Additionally, no analyses were conducted for Depression and Pathological Voice, as Pathological Voice employs a Random Forest model, and Depression is a multi-classification model that does not apply a threshold to its outputs. The results can be seen in, A.7 (Heart Disease), A.8 (Alzheimer Disease), A.9 (Breast Cancer), A.10 (Chronic Kidney Disease), A.12 (Multiple Sclerosis), A.11 (Differentiated Thyroid Cancer Recurrence) and in table 4.3 which lists the optimal threshold values according to the heuristic.

Tabelle 4.3.: The table presents the optimal threshold values, determined using the cutoff heuristic, for both the non-invasive and all features binary classification models.

Dataset	Model Type	Optimal Threshold (Cutoff)
Heart Disease	Non-Invasive Model	0.55
	All Features Model	0.53
Breast Cancer	Non-Invasive Model	0.31
	All Features Model	0.53
Alzheimer	Non-Invasive Model	0.51
	All Features Model	0.63
Chronic Kidney Disease	Non-Invasive Model	0.56
	All Features Model	0.62
Multiple Sclerosis	Non-Invasive Model	0.47
	All Features Model	0.59
Differentiated Thyroid Cancer	Non-Invasive Model	0.50
	All Features Model	0.87

4.5. Network Comparison Results

The DeLong statistic was applied to every dataset except for Pathological Voice and Depression. For all datasets analyzed, the DeLong test indicated that there was no statistically significant difference between the area under the receiver operating characteristic (ROC) curves of the models trained on invasive and non-invasive features. Specifically, the p-values obtained from the tests were not low enough to reject the null hypothesis of no difference in performance.

For the Alzheimer dataset, the p-value was 0.1169, indicating a lack of statistical significance, with an AUC of 0.9138 for the non-invasive model compared to 0.8914 for the all features model. Similarly, in the Breast Cancer dataset, the p-value was 0.7609, showing no significant difference between the AUC of 0.9286 for the all features model and 0.9184 for the non-invasive model. The Heart Disease dataset yielded a p-value of 0.5832, with an AUC of 0.9450 for the non-invasive model versus 0.9276 for the invasive model.

The Chronic Kidney Disease dataset showed a perfect AUC of 1 for both models, with a p-value of 1, indicating no difference. The Multiple Sclerosis dataset had a p-value of 0.9238, with AUCs of 0.8333 for the non-invasive model and 0.8310 for the invasive model. Lastly, in the Differentiated Thyroid Cancer Recurrence dataset, the p-value was 0.2085, with AUCs of 0.8982 for the non-invasive model and 1.0 for the invasive model.

The results suggest that the models utilizing non-invasive features performed similarly to those using invasive features across all datasets analyzed.

Tabelle 4.4.: This table compares the DeLong Test results for non-invasive and all-features models across the binary classification datasets, displaying AUC, Z-Statistic, p-value, and 95% Confidence Interval.

Dataset	Non-Invasive AUC	All Features AUC	Z-Statistic	p-value	95% CI
Alzheimer's	0.9138	0.8914	1.5678	0.1169	[-0.0056, 0.0505]
Breast Cancer	0.9184	0.9286	0.3044	0.7609	[-0.0555, 0.0759]
Heart Disease	0.9450	0.9276	0.5488	0.5832	[-0.0449, 0.0798]
Chronic Kidney Disease	1.0000	1.0000	0.0000	1.0000	[0.0000, 0.0000]
Multiple Sclerosis	0.8333	0.8310	0.0957	0.9238	[-0.0464, 0.0512]
Differentiated Thyroid Cancer Recurrence	0.8982	1.0000	-1.2576	0.2085	[-0.2604, 0.0568]

4.6. Calibration and Probability Interpretation

Calibration was performed for every dataset except for the Pathological Voice dataset. During this process, inconsistencies were encountered while attempting to fit a probability distribution to the class scores. Specifically, for some datasets, such as the Heart Disease dataset, the non-invasive models positive class scores exhibited unrealistic drops in the distribution with the highest likelihood, which was a Cauchy distribution. This led to inaccurate calibrations for certain score values. In the case of scores under 0.25, the calibration consistently returned a value of 1. To resolve this, a Normal distribution was forced for the Heart Disease dataset, which had a more stable fit. While this adjustment did not yield perfectly optimal calibrations, it provided more reliable results compared to the initial approach.

The calibrated model scores were subsequently used in the next phase of analysis, namely the pivot frame validation, which is explained in the following section.

The following tables A.21 and A.22 summarize the calibration errors (ECE, MCE, LCE) for each dataset across both models.

4.7. Pivotframes for Pattern Validation

To summarize, for some diseases, distinct patterns can be seen between features, while for others, visible patterns were less present. For Alzheimer, the strongest indicators for positive predictions were memory complaints combined with older age and low functional assessment scores. In breast cancer, age had a significant impact, with older age consistently leading to positive predictions. For heart disease, features like Oldpeak, severe chest pain, and Exercise Angina showed a clear trend toward predicting the disease. In contrast, models for Chronic Kidney Disease and Multiple Sclerosis showed substantial overfitting to the positive class, making it difficult to observe meaningful patterns. For Pathological Voice, the VHI score and age strongly influenced predictions, while tomatoes eaten per day had no impact. Differentiated Thyroid Cancer Recurrence patterns included higher risk values and later stages correlating with positive predictions. For depression, the CI midpoint

was the most important feature in identifying the people with the condition. Overall, the GUESS-calibrated tables and model predictions aligned closely.

For a complete overview, the pivot tables for their respective dataset can be inspected. Figures A.25, A.26, A.27, A.28 (Heart Disease), figures A.13, A.14, A.15, A.16 (Alzheimer Disease), figures A.17, A.18, A.19, A.20 (Breast Cancer), figures A.21, A.22, A.23, A.24 (Depression), figures A.29, A.30, A.31, A.32 (Chronic Kidney Disease), figures, A.33, A.34, A.35, A.36 (Multiple Sclerosis), figure A.37 (Pathological Voice), and figures A.38, A.39, A.40, A.41 (Differentiated Thyroid Cancer Recurrence).

4.8. Intervention Analysis Based on Blood Pressure

The results, as shown in Table 4.5, reveal a slight increase in mean values for very low and very high blood pressure levels compared to intermediate values. This pattern is consistent across all three analysis methods. Notably, blood pressure values below 90 and above 150 exhibit higher probability and score values, indicating that extremely low or high blood pressure is associated with a marginally higher likelihood of heart disease. However, these values remain relatively low overall.

For older populations, blood pressure below 120 is often considered low, while 130-140 is more typical. Our results align with this, as the lowest model outputs occur around 140, consistent with normal blood pressure in this demographic. Higher blood pressure levels (e.g., above 150) are again associated with increased prediction values, as expected for older age groups.

Tabelle 4.5.: This table compares the mean values of calibrations, scores, and threshold-applied predictions across different systolic blood pressure values, while the other features were generated from a uniform distribution.

Blood Pressure	Guess Calibrated	Model Scores	Threshold Application
bp80	0.6213	0.6562	0.6656
bp100	0.5578	0.6041	0.6002
bp120	0.5248	0.5754	0.5602
bp140	0.5185	0.5687	0.5541
bp160	0.5327	0.5783	0.5703
bp180	0.5562	0.5961	0.5974
bp200	0.5841	0.6168	0.6258

4.9. Causal Discovery on Heart Disease and Diabates Dataset

To identify potential causal relationships in the heart disease dataset, the Fast Causal Inference (FCI) algorithm was applied with five different conditional independence tests on a selected feature subset, including age, sex, oldpeak, exercise-induced angina, resting blood

pressure, the slope of the peak exercise ST segment, and the target variable heart disease. The results for each run are summarized below, with attention to the causal structure suggested by various edge types.

In the first run, using the Fisher's Z test, limited causal insights emerged as seen in A.42. Many relationships between features and the target variable showed ambiguous connections, indicating no clear evidence of causal influence between these variables. However, some edges suggest potential influences: age may impact resting blood pressure, and exercise-induced angina may also influence it. Additionally, age, heart disease, oldpeak, and exercise-induced angina appeared to impact the slope of the peak exercise ST segment, though without confirming direct causality.

In the second run, the Chi-squared test provided a more defined structure, revealing both directed and bidirectional connections A.43. The variable sex showed no associations, indicating limited or no connection with other variables in this subset. Resting blood pressure appeared to influence oldpeak, establishing a directional relationship. Oldpeak seemed to causally influence exercise-induced angina. A bidirectional relationship between oldpeak and the slope of the peak ST segment suggested a latent common cause, hinting at an unknown factor influencing both. There were also indications that age might influence heart disease, while exercise-induced angina appeared to have a direct causal impact on heart disease. Furthermore, a bidirectional connection between heart disease and the slope of the peak ST segment hinted at another latent factor affecting both.

The G-squared test, used in the third run, produced distinct dependencies with a mixture of directed and bidirectional edges A.44. Age and resting blood pressure showed no connections, implying they may not influence other variables in this subset. The data suggested that sex could impact heart disease. Heart disease showed direct relationships with both exercise-induced angina and the slope of the peak ST segment, suggesting causal connections. A bidirectional association between heart disease and oldpeak pointed to a possible latent factor, with oldpeak displaying dependencies with exercise-induced angina, heart disease, and the slope of the peak exercise ST segment, likely due to hidden influences. An ambiguous connection between exercise-induced angina and the slope of the peak exercise ST segment indicated no clear information on causation.

The fourth run, applying the Kernel-Based Conditional Independence (KCI) test, revealed complex relationships with several bidirectional and undirected edges A.45. Age was associated with both resting blood pressure and heart disease without clear causality. Additionally, age had a potential influence on oldpeak, though uncertain. The variable sex showed a unique relationship with heart disease, suggesting a non-standard connection without directional causality. Heart disease had directed relationships with both oldpeak and exercise-induced angina, as well as with the slope of the peak exercise ST segment. Exercise-induced angina showed dependencies with both oldpeak and the slope of the peak ST segment, indicating latent common causes, with oldpeak also directly impacting the slope

of the peak exercise ST segment and displaying mutual dependence with exercise-induced angina.

In the final run, conducted with the Missing Value Fisher's Z test, results were consistent with those from the initial Fisher's Z test, showing limited insights and ambiguous causality across many relationships A.46.

The causal discovery results across all tests reveal significant ambiguities and limitations within this dataset, demonstrated by the inconsistencies in edge direction and strength between each run. Although some associations, like the influence of oldpeak on exercise-induced angina, emerged consistently, many other connections did not align well with known physiological relationships and were open to varying interpretations. For instance, expected relationships, such as the influence of age on blood pressure, were not consistently captured, suggesting limitations in the graph's ability to reflect real-world causality.

Following the analysis on heart disease, a similar causal discovery approach was applied to a diabetes dataset using the Fisher's Z test, visualized in A.47. However, the resulting causal graph for diabetes displayed significant ambiguities that limit its usability for deriving causal insights.

5. Discussion

The results of the feature importance analysis suggest that for almost all datasets, the models can be effectively reduced to non-invasive features without a significant loss in performance. This highlights the potential of non-invasive features to provide accurate predictions, offering a practical alternative to invasive methods in disease prediction. Specifically, neural networks demonstrated the best performance for most datasets, except for the Pathological Voice dataset, where the Random Forest model trained on the top 18 features outperformed the neural network. This is likely due to the insufficient data available for training the convolutional neural network (CNN), which typically requires a larger and more diverse dataset to perform optimally. The results provide valuable insights that can be used in the development of Clinical Decision Support Systems, facilitating more accessible and accurate diagnostics based on non-invasive data.

All models demonstrated strong predictive performance on their respective test datasets. For the chronic kidney disease and differentiated thyroid cancer recurrence datasets, other study groups, such as the Department of Computer Science and Engineering, Manipal University Jaipur, India, and the Department of Endocrinology, Faculty of Medicine, Hamadan University of Medical Sciences, Iran, have conducted similar research [146, 136]. Our models trained on all features consistently outperformed those from the study groups across all evaluated performance metrics. For example, while Chittora et al. chronic kidney disease model achieved a recall of 0.93 and an accuracy of 0.94, our model reached perfect scores with a recall and accuracy of 1.0 [146]. Similarly, for the differentiated thyroid cancer recurrence model, our model achieved both a recall and specificity of 1.0, compared to Borzooei et al. highest results of 0.99 recall and 0.94 specificity [136]. This could be due to differences in the preprocessing stage by Chittora et al. and Borzooei et al. for these datasets [146, 136]. It is important to note that comparing models trained on different data splits can be challenging, and it's not always appropriate to rank one model above another without further investigating actual generalization of the models.

Despite exploring optimal thresholds, no significant improvements in performance were observed, which was expected given that this approach is purely heuristic, aiming to balance the true positive and false positive rates. From a medical perspective, particularly in the context of disease predictions based on non-invasive data, an alternative cutoff approach may be more appropriate. Specifically, employing a precision-recall tradeoff that optimizes for recall—rather than precision—could lead to an increased true positive rate, though at the cost of a higher false positive rate. In the case of non-invasive data, this tradeoff

might be acceptable, as some degree of misclassification could be tolerated if it ensures that individuals with the disease are correctly classified. However, before making this tradeoff, it is important to consult with medical professionals to determine whether a higher true positive rate is actually desirable for the specific disease context. This consultation is crucial because predicting a healthy individual as having the disease will likely lead hospitals to conduct follow-up invasive procedures to confirm the diagnosis. Once there is clarity from medical professionals on this balance, we can consider this tradeoff as a possible approach for specific applications. This is particularly relevant given that non-invasive procedures are generally less costly and less resource-intensive than invasive ones. Therefore, the misclassification of healthy individuals would be less severe, making the benefits of accurate disease classification outweigh the drawbacks of some false positives.

The DeLong test, which compared the models trained on all features and non-invasive features, showed no statistically significant difference between the two approaches for any dataset. This suggests that models based on non-invasive features can deliver sufficient predictions or, at least, indicate a tendency for a condition, further validating the potential for using non-invasive data in medical applications. The lack of a significant difference also reinforces the argument for the feasibility and utility of non-invasive prediction methods in clinical settings where invasive testing may be less accessible or cost-prohibitive. However, it is important to note that invasive procedures cannot be entirely replaced, as they are still necessary for certain tests, such as blood tests, which provide critical diagnostic information. While non-invasive features worked well for these specific datasets, invasive procedures continue to play a crucial role in comprehensive medical assessments.

The pivot tables provided essential insights into how well the neural networks generalized beyond the training data, particularly in cases where traditional performance metrics could have been misleading. For both the Breast Cancer and Chronic Kidney Disease datasets, the models trained on both non-invasive and all features displayed signs of overfitting, while for Multiple Sclerosis, only the all-features model exhibited this issue. In these cases, the pivot tables revealed a strong bias toward the positive class, with very few negative predictions, indicating that the models failed to generalize effectively. This suggests that the models may have learned patterns specific to the dataset, rather than broader generalizable trends. Without the pivot tables, these biases would remain hidden, as the models appeared nearly perfect based on traditional performance metrics but would likely underperform on new data. On the other hand, the pivot tables for other disease models yielded valuable insights into the feature interactions that the neural network learned. This approach helped verify that these models not only achieved balanced predictive power but also learned meaningful patterns rather than simply reflecting dataset biases. However, pivot tables can become complex and difficult to interpret when a lot features are involved, which can limit their usability for high dimensional data.

The intervention analysis indicated that blood pressure has minimal influence compared to other features, which limits its practical impact from a medical perspective. This is

especially significant because we aim to identify features that could lead to actionable interventions. However, the predictions in this analysis appeared to be more closely aligned with patterns in the age feature, as the output showed only slight variation for different blood pressure values. This suggests that the model may have learned an age-related bias, rather than capturing the expected influence of blood pressure. The limited variability in predictions related to blood pressure might also be attributed to the narrow age range within the dataset. A more diverse population with a wider age range might allow for a clearer understanding of how blood pressure independently affects predictions, potentially leading to more meaningful intervention insights. While the direction of predictions in response to varying blood pressure values is correct (lower values for normal blood pressure, higher values for elevated blood pressure), the lack of significant variation reduces the interpretability and clinical applicability of this feature in the current dataset.

The causal discovery process revealed significant limitations due to the bottleneck effect caused by high dimensionality and low sample size, which affected the statistical independence tests within the causal algorithms. Additionally, the presence of mixed data types, for instance metric, categorical or nominal data may have further complicated the CI tests, introducing challenges in accurately assessing relationships between variables. This bottleneck hindered the generation of meaningful causal graphs, as none of the derived causal structures revealed a clear or reliable relationship between specific features and heart disease and diabetes. Consequently, without a dependable causal graph, making well-founded interventions becomes challenging, as we lack insight into the actual causal relationships. For deriving actionable causal insights in a medical context, interventions need to be based on interventional probabilities, $P(y \mid do(x))$, where $do(x)$ represents an active intervention rather than simple observation. To achieve this, do-calculus is required, a formal method that enables the expression of interventions by reducing them to probabilities [147]. However, do-calculus relies on the existence of an accurate causal graph and requires a generative model that estimates the joint probability distribution for all variables involved, to make accurate interventions. In this thesis, however, we used a discriminative approach, which, by design, does not support full causal inference or the derivation of interventional effects.

Ideally, our models output conditional probabilities, represented by the probability of the disease given the input features. However, for a more complete causal understanding, we would ideally compute the joint probability of the probability of the disease given the input features. For the joint probability we need the conditional probability and the probability distribution for the feature configurations.

In our current discriminative model (a neural network), we only try to estimate the probability of the disease given the input features, focusing on learning a decision boundary between disease and non-disease cases. This conditional probability alone limits our ability to simulate realistic interventions since we lack a full model of the data distribution. Generative models learn this joint probability, making them more suitable for causal

inference as they allow us to estimate the probability of the features [148]. This is done by empirically counting how often each combination of features occurs across the total dataset, providing a complete picture of how the features are distributed in the data [148].

While the generated causal graphs provided an initial look at potential dependencies, inconsistencies in the relationships and their lack of alignment with known physiological connections highlighted that these graphs did not capture true causality.

6. Conclusion

In conclusion, this thesis demonstrated that datasets which were reduced to their non-invasive features could be used in disease prediction across various datasets. The developed models, especially the ones for Heart Disease, Alzheimer Disease, Breast Cancer, Multiple Sclerosis, Pathological Voice and Differentiated Thyroid Cancer Recurrence, can be used for Clinical Decision Support Systems and can be implemented in the Virtual Doctor. However, as indicated in the results, this approach may still be limited. The pivot tables show that some models failed to generalize well, not due to the non-invasive feature subset, but due to overfitting on the training data or because of an inherent bias in the dataset. The pivot tables with bad generalization displayed almost only positive predictions. This suggests the model is biased towards predicting the positive class. This poor generalization indicates a bias in the original dataset since the tables were created based on synthetic data. Specifically, when the test set metrics are nearly perfect, including the bias, which may represent a strong pattern in the dataset, the model performs poorly on synthetic data. This suggests that the synthetic data lacks the bias patterns of the original dataset, which exposes the models poor generalization. However, if these models are implemented into the Virtual Doctor, they can be further tested and fine-tuned with newly collected data, which may help reduce the impact of such biases and improve generalization in real-world applications.

The intervention analysis has shown that the lack of a generative model posed a limitation. True intervention analysis requires the joint probabilities. Since this thesis used discriminative models, conditional probabilities can still be used to suggest potential interventions. This approach won't work every time since we hold one feature constant and uniform simulate the rest of the data which doesn't reflect real world scenarios. It requires a dataset which has patterns that can lead to deriving interventions in real world applications. For instance, a high blood pressure might indicate heart disease, which wasn't the case for our heart disease dataset.

Causal discovery was similarly restricted by the high dimensionality and limited sample sizes of the data, making it difficult to extract robust causal insights. Although these challenges limited our causal analyses, they also highlight promising areas for future research, particularly in the development of generative models and in addressing sample-size limitations.

In summary, this thesis advances the exploration of machine learning for non-invasive

diagnostics but reveals key areas for improvement. Addressing data bias, refining feature selection, and incorporating generative approaches could significantly enhance the robustness of non-invasive predictions and enable more reliable causal discoveries in healthcare applications.

In addition, further exploration of use-cases for the Virtual Doctor, which are based on non-invasive features, has yielded usable results. In particular, the inclusion of appropriate sensors is crucial for the integration of these models. For pathological voice, a basic voice capturing microphone is recommended, while for heart disease, sensors for blood pressure, like blood pressure monitors, and heart rate, like photoplethysmography would be necessary.

Additionally, the ability to derive interventions through neural networks was explored. While further work is needed to understand the distribution and relationships between features, early results with uniformly Monte Carlo simulated features suggest that such interventions could be usable and extendable.

While causal discovery methods showed limitations in this study, it was noted that neural networks often outperform statistical approaches, for instance causal discovery in terms of pattern recognition and generalization, making them more suitable for reliable predictions in many healthcare contexts.

Finally, the datasets inherent bias was discussed, highlighting the impact of such bias on model performance and generalization.

Literatur

[1] World Health Organization. *Health and Care Workforce: Addressing Challenges in Rural and Remote Areas*. Accessed: 2024-10-29. 2021. URL: https://www.who.int/campaigns/annual-theme/year-of-health-and-care-workers-2021/facts.

[2] World Health Organization. *Retention of the Health Workforce in Rural and Remote Areas: A Systematic Review*. Accessed: 2024-10-29. 2020. URL: https://www.who.int/news/item/08-12-2020-retention-of-the-health-workforce-in-rural-and-remote-areas-a-systematic-review.

[3] World Health Organization. *Health Workforce Development and Retention in Rural Areas*. Accessed: 2024-10-29. 2023. URL: https://www.who.int/teams/health-workforce/health-workforce-development/.

[4] James Buchan u. a. „Global implications of health workforce shortages and the role of policymakers“. In: *Lancet Global Health* 11.3 (2023). Discusses the influence of healthcare staff shortages on patient care quality., e333–e339.

[5] Ifeanyi Anthony Okwor u. a. „Digital Technologies Impact on Healthcare Delivery: A Systematic Review of Artificial Intelligence (AI) and Machine-Learning (ML) Adoption, Challenges, and Opportunities“. In: *AI* 5.4 (2024), S. 1918–1941. ISSN: 2673-2688. DOI: 10.3390/ai5040095. URL: https://www.mdpi.com/2673-2688/5/4/95.

[6] Oliver Higgins u. a. „Artificial intelligence (AI) and machine learning (ML) based decision support systems in mental health: An integrative review“. In: *International Journal of Mental Health Nursing* 32.4 (2023), S. 966–978. DOI: https://doi.org/10.1111/inm.13114. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/inm.13114. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/inm.13114.

[7] Molly Bekbolatova u. a. „Transformative Potential of AI in Healthcare: Definitions, Applications, and Navigating the Ethical Landscape and Public Perspectives“. en. In: *Healthcare (Basel)* 12.2 (Jan. 2024).

[8] Mohamed Said Ibrahim und Sameh Saber. „Machine Learning and Predictive Analytics: Advancing Disease Prevention in Healthcare“. In: 7 (Jan. 2023), S. 53–71. URL: https://publications.dlpress.org/index.php/jcha/article/view/16.

[9] Lorenz Adlung u. a. „Machine learning in clinical decision making“. In: *Med* 2.6 (2021), S. 642–665. ISSN: 2666-6340. DOI: https://doi.org/10.1016/j.medj.2021.04.006. URL: https://www.sciencedirect.com/science/article/pii/S2666634021001550.

[10] Sebastian Spänig u. a. „The virtual doctor: An interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes“. In: *Artificial Intelligence in Medicine* 100 (2019), S. 101706. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2019.101706. URL: https://www.sciencedirect.com/science/article/pii/S0933365719301083.

[11] Joris L J M Müskens u. a. „Overuse of diagnostic testing in healthcare: a systematic review“. In: *BMJ Quality & Safety* 31.1 (2022), S. 54–63. ISSN: 2044-5415. DOI: 10.1136/bmjqs-2020-012576. eprint: https://qualitysafety.bmj.com/content/31/1/54.full.pdf. URL: https://qualitysafety.bmj.com/content/31/1/54.

[12] Said A. Ibrahim u. a. „Diagnostic Errors, Health Disparities, and Artificial Intelligence: A Combination for Health or Harm?“ In: *JAMA Health Forum* 2.9 (2021), e213243. DOI: 10.1001/jamahealthforum.2021.2430.

[13] William Weeks u. a. „Rural-urban disparities in health outcomes, clinical care, health behaviors, and social determinants of health and an action-oriented, dynamic tool for visualizing them“. In: *PLOS Global Public Health* 3 (Okt. 2023), e0002420. DOI: 10.1371/journal.pgph.0002420.

[14] Ariya Natarajan u. a. „Access to Healthcare in Rural Communities: A Bibliometric Analysis“. en. In: *Health Psychol Res* 11 (Dez. 2023), S. 90615.

[15] Andrew Henderson. „Coronary heart disease: Overview“. In: *The Lancet* 348 (Nov. 1996), S1–S4. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(96)98001-0. URL: https://doi.org/10.1016/S0140-6736(96)98001-0.

[16] Manish Shah, Fadi G. Akar und Gordon F. Tomaselli. „Molecular Basis of Arrhythmias“. In: *Circulation* 112.16 (2005), S. 2517–2529. DOI: 10.1161/CIRCULATIONAHA.104.494476. eprint: https://www.ahajournals.org/doi/pdf/10.1161/CIRCULATIONAHA.104.494476. URL: https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.104.494476.

[17] Sonali S. Patel und Trudy L. Burns. „Nongenetic Risk Factors and Congenital Heart Defects“. In: *Pediatric Cardiology* 34.7 (Okt. 2013), S. 1535–1555. ISSN: 1432-1971. DOI: 10.1007/s00246-013-0775-4. URL: https://doi.org/10.1007/s00246-013-0775-4.

[18] World Health Organization. *Cardiovascular Diseases*. https://www.who.int/health-topics/cardiovascular-diseases (Accessed: 2024-10-30). 2024.

[19] Mariachiara Di Cesare u. a. „The Heart of the World“. en. In: *Glob Heart* 19.1 (Jan. 2024), S. 11.

[20] Sam Richmond und Christopher Wren. „Early diagnosis of congenital heart disease“. In: *Seminars in Neonatology* 6.1 (2001), S. 27–35. ISSN: 1084-2756. DOI: https://doi.org/10.1053/siny.2000.0028. URL: https://www.sciencedirect.com/science/article/pii/S1084275600900289.

[21] Carlos Collet u. a. „State of the art: coronary angiography“. In: *EuroIntervention* 13.6 (2017), S. 634–643. DOI: 10.4244/EIJ-D-17-00465. URL: https://doi.org/10.4244/EIJ-D-17-00465.

[22] Ana Mincholé u. a. „Machine learning in the electrocardiogram“. In: *Journal of Electrocardiology* 57 (2019), S61–S64. ISSN: 0022-0736. DOI: https://doi.org/10.1016/j.jelectrocard.2019.08.008. URL: https://www.sciencedirect.com/science/article/pii/S0022073619304571.

[23] Charulatha Ramanathan u. a. „Noninvasive electrocardiographic imaging for cardiac electrophysiology and arrhythmia“. In: *Nature Medicine* 10.4 (Apr. 2004), S. 422–428. ISSN: 1546-170X. DOI: 10.1038/nm1011. URL: https://doi.org/10.1038/nm1011.

[24] Matthias Christgen u. a. „Lobular breast cancer: Clinical, molecular and morphological characteristics“. In: *Pathology - Research and Practice* 212.7 (2016), S. 583–597. ISSN: 0344-0338. DOI: https://doi.org/10.1016/j.prp.2016.05.002. URL: https://www.sciencedirect.com/science/article/pii/S0344033816300826.

[25] William E. Bentley u. a. „Multiple signaling pathways contribute to the induction of cyclooxygenase-2 expression in breast cancer cells“. In: *The Journal of Clinical Investigation* 117.11 (2007), S. 3155–3163. DOI: 10.1172/JCI33295. URL: https://www.jci.org/articles/view/33295.

[26] Melina Arnold u. a. „Current and future burden of breast cancer: Global statistics for 2020 and 2040“. en. In: *Breast* 66 (Dez. 2022), S. 15–23.

[27] Ophira Ginsburg u. a. „Breast cancer early detection: A phased approach to implementation“. en. In: *Cancer* 126 Suppl 10.Suppl 10 (Mai 2020), S. 2379–2393.

[28] N. Houssami u. a. „Early detection of second breast cancers improves prognosis in breast cancer survivors“. In: *Annals of Oncology* 20.9 (2009), S. 1505–1510. ISSN: 0923-7534. DOI: https://doi.org/10.1093/annonc/mdp037. URL: https://www.sciencedirect.com/science/article/pii/S0923753419404468.

[29] Robert A. Smith. „An Overview of Mammography: Benefits and Limitations“. In: *Journal of the National Comprehensive Cancer Network* 1.2 (2003), S. 264–271. DOI: 10.6004/jnccn.2003.0024. URL: https://jnccn.org/view/journals/jnccn/1/2/article-p264.xml.

[30] Hong Zhao u. a. „Limitations of mammography in the diagnosis of breast diseases compared with ultrasonography: a single-center retrospective analysis of 274 cases“. In: *European Journal of Medical Research* 20.1 (Apr. 2015), S. 49. ISSN: 2047-783X. DOI: 10.1186/s40001-015-0140-6. URL: https://doi.org/10.1186/s40001-015-0140-6.

[31] Arslan Khalid u. a. „Breast Cancer Detection and Prevention Using Machine Learning“. In: *Diagnostics* 13.19 (2023). ISSN: 2075-4418. DOI: 10.3390/diagnostics13193113. URL: https://www.mdpi.com/2075-4418/13/19/3113.

[32] Zeinab Breijyeh und Rafik Karaman. „Comprehensive Review on Alzheimer's Disease: Causes and Treatment“. In: *Molecules* 25.24 (2020). ISSN: 1420-3049. DOI: 10.3390/molecules25245789. URL: https://www.mdpi.com/1420-3049/25/24/5789.

[33] Jose A. Soria Lopez, Hector M. González und Gabriel C. Léger. „Chapter 13 - Alzheimer's disease“. In: *Geriatric Neurology*. Hrsg. von Steven T. Dekosky und Sanjay Asthana. Bd. 167. Handbook of Clinical Neurology. Elsevier, 2019, S. 231–255. DOI: https://doi.org/10.1016/B978-0-12-804766-8.00013-3. URL: https://www.sciencedirect.com/science/article/pii/B9780128047668000133.

[34] Harald Hampel u. a. „The use of lumbar puncture and safety recommendations in Alzheimer's disease: a plain language summary“. en. In: *Neurodegener Dis Manag* 12.5 (Juli 2022), S. 221–229.

[35] J. D. Warren u. a. „Brain biopsy in dementia“. In: *Brain* 128.9 (Mai 2005), S. 2016–2025. ISSN: 0006-8950. DOI: 10.1093/brain/awh543. eprint: https://academic.oup.com/brain/article-pdf/128/9/2016/1044537/awh543.pdf. URL: https://doi.org/10.1093/brain/awh543.

[36] Yi-Ting Tina Wang, Pedro Rosa-Neto und Serge Gauthier. „Advanced brain imaging for the diagnosis of Alzheimer disease“. In: *Current Opinion in Neurology* 36.5 (2023). URL: https://journals.lww.com/co-neurology/fulltext/2023/10000/advanced_brain_imaging_for_the_diagnosis_of.19.aspx.

[37] Kurt A Jellinger u. a. „Biomarkers for early diagnosis of Alzheimer disease: 'ALZheimer ASsociated gene'–a new blood biomarker?“ en. In: *J Cell Mol Med* 12.4 (März 2008), S. 1094–1117.

[38] Sonja Simpraga u. a. „EEG machine learning for accurate detection of cholinergic intervention and Alzheimer's disease“. In: *Scientific Reports* 7.1 (Juli 2017), S. 5775. ISSN: 2045-2322. DOI: 10.1038/s41598-017-06165-4. URL: https://doi.org/10.1038/s41598-017-06165-4.

[39] Paolo Cassano und Maurizio Fava. „Depression and public health: An overview“. In: *Journal of Psychosomatic Research* 53.4 (2002), S. 849–857. ISSN: 0022-3999. DOI: https://doi.org/10.1016/S0022-3999(02)00304-5. URL: https://www.sciencedirect.com/science/article/pii/S0022399902003045.

[40] Danuta Wasserman. *Depression*. OUP Oxford, 2011.

[41] Pim Cuijpers u. a. „Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: A meta-analysis“. In: *Clinical Psychology Review* 30.6 (2010), S. 768–778. ISSN: 0272-7358. DOI: https://doi.org/10.1016/j.cpr.2010.06.001. URL: https://www.sciencedirect.com/science/article/pii/S0272735810000954.

[42] World Health Organization. *Depressive disorder (depression)*. Accessed: 2024-10-31. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/depression.

[43] Smitha Joyce Pinto und Mimmo Parente. „Comprehensive review of depression detection techniques based on machine learning approach“. In: *Soft Computing* 28.17 (Sep. 2024), S. 10701–10725. ISSN: 1433-7479. DOI: 10.1007/s00500-024-09862-1. URL: https://doi.org/10.1007/s00500-024-09862-1.

[44] Sana Yasin u. a. „Machine learning based approaches for clinical and non-clinical depression recognition and depression relapse prediction using audiovisual and EEG modalities: A comprehensive review“. In: *Computers in Biology and Medicine* 159 (2023), S. 106741. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2023.106741. URL: https://www.sciencedirect.com/science/article/pii/S0010482523002068.

[45] Tashfia Ahmed, Meha Qassem und Panicos A. Kyriacou. „Physiological monitoring of stress and major depression: A review of the current monitoring techniques and considerations for the future“. In: *Biomedical Signal Processing and Control* 75 (2022), S. 103591. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2022.103591. URL: https://www.sciencedirect.com/science/article/pii/S1746809422001136.

[46] Kwang-Sig Lee und Byung-Joo Ham. „Machine Learning on Early Diagnosis of Depression“. en. In: *Psychiatry Investig* 19.8 (Aug. 2022), S. 597–605.

[47] Csaba P Kovesdy. „Epidemiology of chronic kidney disease: an update 2022“. en. In: *Kidney Int Suppl (2011)* 12.1 (März 2022), S. 7–11.

[48] Angela C. Webster u. a. „Chronic Kidney Disease“. In: *The Lancet* 389.10075 (März 2017), S. 1238–1252. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(16)32064-5. URL: https://doi.org/10.1016/S0140-6736(16)32064-5.

[49] Pablo Ureña Torres u. a. „When, How, and Why a Bone Biopsy Should Be Performed in Patients With Chronic Kidney Disease“. In: *Seminars in Nephrology* 34.6 (2014). Mineral Bone Disorder Is a Key Player in Chronic Kidney Disease, S. 612–625. ISSN: 0270-9295. DOI: https://doi.org/10.1016/j.semnephrol.2014.09.004. URL: https://www.sciencedirect.com/science/article/pii/S0270929514001442.

[50] Justyna Siwy u. a. „Noninvasive diagnosis of chronic kidney diseases using urinary proteome analysis“. In: *Nephrology Dialysis Transplantation* 32.12 (Okt. 2016), S. 2079–2089. ISSN: 0931-0509. DOI: 10.1093/ndt/gfw337. eprint: https://academic.oup.com/ndt/article-pdf/32/12/2079/24363496/gfw337.pdf. URL: https://doi.org/10.1093/ndt/gfw337.

[51] Nazem Ghasemi, Shahnaz Razavi und Elham Nikzad. „Multiple Sclerosis: Pathogenesis, Symptoms, Diagnoses and Cell-Based Therapy“. en. In: *Cell J* 19.1 (Dez. 2016), S. 1–10.

[52] Clare Walton u. a. „Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition“. en. In: *Mult Scler* 26.14 (Nov. 2020), S. 1816–1821.

[53] So Yamada u. a. „Tumefactive multiple sclerosis requiring emergent biopsy and histological investigation to confirm the diagnosis: a case report“. In: *Journal of Medical Case Reports* 6 (2012), S. 1–4.

[54] M.D. KEITH I. MARTON und M.D. ALISA D. GEAN. „The Spinal Tap: A New Look at an Old Test“. In: *Annals of Internal Medicine* 104.6 (1986). PMID: 3518565, S. 840–848. DOI: 10.7326/0003-4819-104-6-840. eprint: https://www.acpjournals.org/doi/pdf/10.7326/0003-4819-104-6-840. URL: https://www.acpjournals.org/doi/abs/10.7326/0003-4819-104-6-840.

[55] Moses Rodriguez und Bernd Scheithauer. „Ultrastructure of Multiple Sclerosis“. In: *Ultrastructural Pathology* 18.1-2 (1994). PMID: 8191643, S. 3–13. DOI: 10.3109/01913129409016267. eprint: https://doi.org/10.3109/01913129409016267. URL: https://doi.org/10.3109/01913129409016267.

[56] Rohit Bakshi u. a. „MRI in multiple sclerosis: current status and future prospects“. In: *The Lancet Neurology* 7.7 (Juli 2008), S. 615–625. ISSN: 1474-4422. DOI: 10.1016/S1474-4422(08)70137-6. URL: https://doi.org/10.1016/S1474-4422(08)70137-6.

[57] Valentina Tomassini u. a. „Diagnosis and management of multiple sclerosis: MRI in clinical practice“. In: *Journal of Neurology* 267.10 (Okt. 2020), S. 2917–2925. ISSN: 1432-1459. DOI: 10.1007/s00415-020-09930-0. URL: https://doi.org/10.1007/s00415-020-09930-0.

[58] William R. Burns und Martha A. Zeiger. „Differentiated Thyroid Cancer“. In: *Seminars in Oncology* 37.6 (2010). Endocrine Cancers, S. 557–566. ISSN: 0093-7754. DOI: https://doi.org/10.1053/j.seminoncol.2010.10.008. URL: https://www.sciencedirect.com/science/article/pii/S0093775410001788.

[59] Ashok R. Shaha. „Recurrent Differentiated Thyroid Cancer“. In: *Endocrine Practice* 18.4 (2012), S. 600–603. ISSN: 1530-891X. DOI: https://doi.org/10.4158/EP12047.CO. URL: https://www.sciencedirect.com/science/article/pii/S1530891X20430666.

[60] Andrés Coca-Pelaz u. a. „Recurrent Differentiated Thyroid Cancer: The Current Treatment Options“. en. In: *Cancers (Basel)* 15.10 (Mai 2023).

[61] Nathan A. Johnson, Shane O. LeBeau und Mitchell E. Tublin. „Imaging Surveillance of Differentiated Thyroid Cancer“. In: *Radiologic Clinics* 49.3 (Mai 2011), S. 473–487. ISSN: 0033-8389. DOI: 10.1016/j.rcl.2011.02.008. URL: https://doi.org/10.1016/j.rcl.2011.02.008.

[62] Jung Hwan Baek u. a. „Locoregional Control of Metastatic Well-Differentiated Thyroid Cancer by Ultrasound-Guided Radiofrequency Ablation“. In: *American Journal of Roentgenology* 197.2 (2011). PMID: 21785061, W331–W336. DOI: 10.2214/AJR.10.5345. eprint: https://doi.org/10.2214/AJR.10.5345. URL: https://doi.org/10.2214/AJR.10.5345.

[63] ANDREA FRASOLDATI u. a. „Role of Thyroglobulin Measurement in Fine-Needle Aspiration Biopsies of Cervical Lymph Nodes in Patients with Differentiated Thyroid Cancer“. In: *Thyroid®* 9.2 (1999). PMID: 10090308, S. 105–111. DOI: 10.1089/thy.1999.9.105. eprint: https://doi.org/10.1089/thy.1999.9.105. URL: https://doi.org/10.1089/thy.1999.9.105.

[64] Tracy S Wang und Julie Ann Sosa. „Thyroid surgery for differentiated thyroid cancer—recent advances and future directions“. In: *Nature Reviews Endocrinology* 14.11 (2018), S. 670–683.

[65] Kwang-Sig Lee und Hyuntae Park. *Machine learning on thyroid disease: a review.* März 2022.

[66] Carole A. Spencer und Jonathan S. LoPresti. „Technology Insight: measuring thyroglobulin and thyroglobulin autoantibody in patients with differentiated thyroid cancer“. In: *Nature Clinical Practice Endocrinology & Metabolism* 4.4 (Apr. 2008), S. 223–233. ISSN: 1745-8374. DOI: 10.1038/ncpendmet0757. URL: https://doi.org/10.1038/ncpendmet0757.

[67] Everthon S. Fonseca und Jose C. Pereira. „Normal versus pathological voice signals“. In: *IEEE Engineering in Medicine and Biology Magazine* 28.5 (2009), S. 44–48. DOI: 10.1109/MEMB.2009.934248.

[68] Lucian Sulica. „Laryngoscopy, stroboscopy and other tools for the evaluation of voice disorders“. In: *Otolaryngol Clin North Am* 46.1 (2013), S. 21–30.

[69] Alireza A Dibazar, Theodore W Berger und Shrikanth S Narayanan. „Pathological voice assessment“. In: *2006 international conference of the IEEE engineering in medicine and biology society.* IEEE. 2006, S. 1669–1673.

[70] Minoru Hirano u. a. „Acoustic analysis of pathological voice: some results of clinical application“. In: *Acta oto-laryngologica* 105.5-6 (1988), S. 432–438.

[71] Vijay Parsa und Donald G. Jamieson. „Acoustic Discrimination of Pathological Voice“. In: *Journal of Speech, Language, and Hearing Research* 44.2 (2001), S. 327–339. DOI: 10.1044/1092-4388(2001/027). eprint: https://pubs.asha.org/doi/pdf/10.1044/1092-4388%282001/027%29. URL: https://pubs.asha.org/doi/abs/10.1044/1092-4388%282001/027%29.

[72] Laura Verde, Giuseppe De Pietro und Giovanna Sannino. „Voice Disorder Identification by Using Machine Learning Techniques“. In: *IEEE Access* 6 (2018), S. 16246–16255. DOI: 10.1109/ACCESS.2018.2816338.

[73] Fahad Taha AL-Dhief u. a. „Voice Pathology Detection Using Machine Learning Technique“. In: *2020 IEEE 5th International Symposium on Telecommunication Technologies (ISTT)*. 2020, S. 99–104. DOI: 10.1109/ISTT50966.2020.9279346.

[74] Heekyu Kim u. a. „Non-invasive way to diagnose dysphagia by training deep learning model with voice spectrograms“. In: *Biomedical Signal Processing and Control* 86 (2023), S. 105259. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2023.105259. URL: https://www.sciencedirect.com/science/article/pii/S1746809423006924.

[75] Mark A. Atkinson und Noel K. Maclaren. „What Causes Diabetes?“ In: *Scientific American* 263.1 (1990), S. 62–71. ISSN: 00368733, 19467087. URL: http://www.jstor.org/stable/24996864 (besucht am 07. 11. 2024).

[76] Juan J. Chillarón u. a. „Type 1 diabetes, metabolic syndrome and cardiovascular risk“. In: *Metabolism* 63.2 (2014), S. 181–187. ISSN: 0026-0495. DOI: https://doi.org/10.1016/j.metabol.2013.10.002. URL: https://www.sciencedirect.com/science/article/pii/S002604951300334X.

[77] Fernando Ovalle und Ricardo Azziz. „Insulin resistance, polycystic ovary syndrome, and type 2 diabetes mellitus“. In: *Fertility and Sterility* 77.6 (2002), S. 1095–1105. ISSN: 0015-0282. DOI: https://doi.org/10.1016/S0015-0282(02)03111-4. URL: https://www.sciencedirect.com/science/article/pii/S0015028202031114.

[78] Merlin C Thomas u. a. „Diabetic kidney disease“. In: *Nature reviews Disease primers* 1.1 (2015), S. 1–20.

[79] Thomas H Marwick. „Diabetic heart disease“. In: *Postgraduate medical journal* 84.990 (2008), S. 188–192.

[80] Ryan Lee, Tien Y Wong und Charumathi Sabanayagam. „Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss“. In: *Eye and vision* 2 (2015), S. 1–25.

[81] Pouya Saeedi u. a. „Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition“. In: *Diabetes Research and Clinical Practice* 157 (2019), S. 107843. ISSN: 0168-8227. DOI: https://doi.org/10.1016/j.diabres.

2019.107843. URL: https://www.sciencedirect.com/science/article/pii/S0168822719312306.

[82] Andre Pascal Kengne u. a. „Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models“. In: *The lancet Diabetes & endocrinology* 2.1 (2014), S. 19–29.

[83] Danyal Mahmood, Hannan Naseem Riaz und Humaira Nisar. „Introduction to Non-Invasive Biomedical Signals for Healthcare“. In: *Advances in Non-Invasive Biomedical Signal Sensing and Processing with Machine Learning*. Springer, 2023, S. 1–24.

[84] Jair Cervantes u. a. „A comprehensive survey on support vector machine classification: Applications, challenges and trends“. In: *Neurocomputing* 408 (2020), S. 189–215. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2019.10.118. URL: https://www.sciencedirect.com/science/article/pii/S0925231220307153.

[85] Shiyu Ji. *Kernel Trick Visualization*. Own work, CC BY-SA 4.0. 2017. URL: https://commons.wikimedia.org/w/index.php?curid=60458994.

[86] Leif Peterson. „K-nearest neighbor“. In: *Scholarpedia* 4 (Jan. 2009), S. 1883. DOI: 10.4249/scholarpedia.1883.

[87] Lior Rokach und Oded Maimon. „Decision Trees“. In: *Data Mining and Knowledge Discovery Handbook*. Hrsg. von Oded Maimon und Lior Rokach. Boston, MA: Springer US, 2005, S. 165–192. ISBN: 978-0-387-25465-4. DOI: 10.1007/0-387-25465-X_9. URL: https://doi.org/10.1007/0-387-25465-X_9.

[88] Leo Breiman. „Random Forests“. In: *Machine Learning* 45.1 (Okt. 2001), S. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

[89] Tianqi Chen und Carlos Guestrin. „XGBoost: A Scalable Tree Boosting System“. In: *CoRR* abs/1603.02754 (2016). arXiv: 1603.02754. URL: http://arxiv.org/abs/1603.02754.

[90] Alexey Natekin und Alois Knoll. „Gradient boosting machines, a tutorial“. In: *Frontiers in Neurorobotics* 7 (2013). ISSN: 1662-5218. DOI: 10.3389/fnbot.2013.00021. URL: https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2013.00021.

[91] Erik Štrumbelj und Igor Kononenko. „Explaining prediction models and individual predictions with feature contributions“. In: *Knowledge and information systems* 41 (2014), S. 647–665.

[92] Bernhard Mehlig. „Artificial Neural Networks“. In: *CoRR* abs/1901.05639 (2019). arXiv: 1901.05639. URL: http://arxiv.org/abs/1901.05639.

[93] Johannes Schmidt-Hieber. „Nonparametric regression using deep neural networks with ReLU activation function“. In: *The Annals of Statistics* 48.4 (2020), S. 1875–1897. DOI: 10.1214/19-AOS1875. URL: https://doi.org/10.1214/19-AOS1875.

[94] Shiv Ram Dubey, Satish Kumar Singh und Bidyut Baran Chaudhuri. „A Comprehensive Survey and Performance Analysis of Activation Functions in Deep Learning“. In: *CoRR* abs/2109.14545 (2021). arXiv: 2109.14545. URL: https://arxiv.org/abs/2109.14545.

[95] Usha Ruby und Vamsidhar Yendapalli. „Binary cross entropy with deep learning technique for image classification“. In: *Int. J. Adv. Trends Comput. Sci. Eng* 9.10 (2020).

[96] Sebastian Ruder. „An overview of gradient descent optimization algorithms“. In: *CoRR* abs/1609.04747 (2016). arXiv: 1609.04747. URL: http://arxiv.org/abs/1609.04747.

[97] Diederik P. Kingma und Jimmy Ba. „Adam: A Method for Stochastic Optimization“. In: *arXiv e-prints*, arXiv:1412.6980 (Dez. 2014), arXiv:1412.6980. DOI: 10.48550/arXiv.1412.6980. arXiv: 1412.6980 [cs.LG].

[98] Ian Goodfellow, Yoshua Bengio und Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[99] Min Xu u. a. „HMM-Based Audio Keyword Generation“. In: *Advances in Multimedia Information Processing - PCM 2004*. Hrsg. von Kiyoharu Aizawa, Yuichi Nakamura und Shin'ichi Satoh. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, S. 566–574. ISBN: 978-3-540-30543-9.

[100] Md. Sahidullah und Goutam Saha. „Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition“. In: *Speech Communication* 54.4 (2012), S. 543–565. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2011.11.004. URL: https://www.sciencedirect.com/science/article/pii/S0167639311001622.

[101] Assim Ara Abdulsatar u. a. „Age and gender recognition from speech signals“. In: *Journal of Physics: Conference Series* 1410.1 (Dez. 2019), S. 012073. DOI: 10.1088/1742-6596/1410/1/012073. URL: https://dx.doi.org/10.1088/1742-6596/1410/1/012073.

[102] Fang Zheng, Guoliang Zhang und Zhanjiang Song. „Comparison of different implementations of MFCC“. In: *Journal of Computer Science and Technology* 16.6 (Nov. 2001), S. 582–589. ISSN: 1860-4749. DOI: 10.1007/BF02943243. URL: https://doi.org/10.1007/BF02943243.

[103] S. Furui. „Speaker-independent isolated word recognition based on emphasized spectral dynamics“. In: *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Bd. 11. 1986, S. 1991–1994. DOI: 10.1109/ICASSP.1986.1168654.

[104] Todor Ganchev, Nikos Fakotakis und Kokkinakis George. „Comparative evaluation of various MFCC implementations on the speaker verification task“. In: *Proceedings of the SPECOM* 1 (Jan. 2005).

[105] Meinard Müller. *Information Retrieval for Music and Motion*. 1. Aufl. Copyright Springer-Verlag Berlin Heidelberg 2007. Springer Berlin, Heidelberg, 2007, S. XVI, 318. ISBN: 978-3-540-74047-6. DOI: 10.1007/978-3-540-74048-3. URL: https://doi.org/10.1007/978-3-540-74048-3.

[106] Kevin W. Bowyer u. a. „SMOTE: Synthetic Minority Over-sampling Technique“. In: *CoRR* abs/1106.1813 (2011). arXiv: 1106.1813. URL: http://arxiv.org/abs/1106.1813.

[107] Haibo He u. a. „ADASYN: Adaptive synthetic sampling approach for imbalanced learning“. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008, S. 1322–1328. DOI: 10.1109/IJCNN.2008.4633969.

[108] Jacqueline Beinecke und Dominik Heider. „Gaussian noise up-sampling is better suited than SMOTE and ADASYN for clinical decision making“. In: *BioData Mining* 14.49 (Nov. 2021). ISSN: 1756-0381. DOI: 10.1186/s13040-021-00283-6. URL: https://doi.org/10.1186/s13040-021-00283-6.

[109] Erwin Otto Kreyszig. *Advanced Engineering Mathematics*. 4th. eq. 5. Wiley, 1979, S. 880. ISBN: 0-471-02140-7.

[110] S. Gopal Krishna Patro und Kishore Kumar Sahu. „Normalization: A Preprocessing Stage“. In: *CoRR* abs/1503.06462 (2015). arXiv: 1503.06462. URL: http://arxiv.org/abs/1503.06462.

[111] Sergey Ioffe und Christian Szegedy. „Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift“. In: *arXiv e-prints*, arXiv:1502.03167 (Feb. 2015), arXiv:1502.03167. DOI: 10.48550/arXiv.1502.03167. arXiv: 1502.03167 [cs.LG].

[112] Jimmy Lei Ba, Jamie Ryan Kiros und Geoffrey E. Hinton. „Layer Normalization“. In: *arXiv e-prints*, arXiv:1607.06450 (Juli 2016), arXiv:1607.06450. DOI: 10.48550/arXiv.1607.06450. arXiv: 1607.06450 [stat.ML].

[113] Jan Benedikt Ruhland, Iraj Masoudian und Dominik Heider. „Enhancing Deep Neural Network Training Through Learnable Adaptive Normalization: Methods and Applications“. Unpublished manuscript, last accessed in November 2024.

[114] Ivo Düntsch und Günther Gediga. „Confusion matrices and rough set data analysis“. In: *CoRR* abs/1902.01487 (2019). arXiv: 1902.01487. URL: http://arxiv.org/abs/1902.01487.

[115] Bradley J Erickson und Felipe Kitamura. *Magician's corner: 9. Performance metrics for machine learning models*. 2021.

[116] Nathalie Japkowicz. „Assessment Metrics for Imbalanced Learning“. In: *Imbalanced Learning*. John Wiley & Sons, Ltd, 2013. Kap. 8, S. 187–206. ISBN: 9781118646106. DOI: https://doi.org/10.1002/9781118646106.ch8. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118646106.ch8. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118646106.ch8.

[117] Andrew P Bradley. „The use of the area under the ROC curve in the evaluation of machine learning algorithms“. In: *Pattern recognition* 30.7 (1997), S. 1145–1159.

[118] E R DeLong, D M DeLong und D L Clarke-Pearson. „Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach“. en. In: *Biometrics* 44.3 (Sep. 1988), S. 837–845.

[119] Johanna Schwarz und Dominik Heider. „GUESS: Projecting Machine Learning Scores to Well-Calibrated Probability Estimates for Clinical Decision Making“. In: *Bioinformatics (Oxford, England)* 35 (Nov. 2018). DOI: 10.1093/bioinformatics/bty984.

[120] Jiji Zhang. „On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias“. In: *Artificial Intelligence* 172.16 (2008), S. 1873–1896. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2008.08.001. URL: https://www.sciencedirect.com/science/article/pii/S0004370208001008.

[121] R. A. Fisher. „Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population“. In: *Biometrika* 10.4 (1915), S. 507–521. ISSN: 00063444. URL: http://www.jstor.org/stable/2331838 (besucht am 23. 10. 2024).

[122] Karl Pearson. „On the Probable Error of a Coefficient of Correlation as Found from a Fourfold Table“. In: *Biometrika* 9.1/2 (1913), S. 22–33. ISSN: 00063444, 14643510. URL: http://www.jstor.org/stable/2331798 (besucht am 23. 10. 2024).

[123] Richa Singhal und Rakesh Rana. „Chi-square test and its application in hypothesis testing“. In: *Journal of the Practice of Cardiovascular Sciences* 1 (Jan. 2015). DOI: 10.4103/2395-5414.157577.

[124] Jesse Hoey. *The Two-Way Likelihood Ratio (G) Test and Comparison to Two-Way Chi Squared Test*. 2012. arXiv: 1206.4881 [stat.ME]. URL: https://arxiv.org/abs/1206.4881.

[125] Kun Zhang u. a. „Kernel-based Conditional Independence Test and Application in Causal Discovery“. In: *CoRR* abs/1202.3775 (2012). arXiv: 1202.3775. URL: http://arxiv.org/abs/1202.3775.

[126] Yujia Zheng u. a. „Causal-learn: Causal discovery in python“. In: *Journal of Machine Learning Research* 25.60 (2024), S. 1–8.

[127] Jiji Zhang. „Causal Reasoning with Ancestral Graphs“. In: *Journal of Machine Learning Research* 9.47 (2008), S. 1437–1474. URL: http://jmlr.org/papers/v9/zhang08a.html.

[128] Shaohua Fan u. a. *Directed Acyclic Graph Structure Learning from Dynamic Graphs*. 2024. arXiv: 2211.17029 [cs.LG]. URL: https://arxiv.org/abs/2211.17029.

[129] Fedesoriano. *Heart Failure Prediction Dataset*. Retrieved [Date Retrieved]. Sep. 2021. URL: https://www.kaggle.com/fedesoriano/heart-failure-prediction.

[130] John Allen. „Photoplethysmography and its application in clinical physiological measurement“. In: *Physiological measurement* 28.3 (2007), R1.

[131] Idongesit Eteng u. a. *UCTH Breast Cancer Dataset*. Version V2. 2023. DOI: 10.17632/63fpbc9cm4.2. URL: https://doi.org/10.17632/63fpbc9cm4.2.

[132] Rabie El Kharoua. *Alzheimer's Disease Dataset*. 2024. DOI: 10.34740/KAGGLE/DSV/8668279. URL: https://www.kaggle.com/dsv/8668279.

[133] Centers for Disease Control und Prevention (CDC). *Indicators of Anxiety or Depression Based on Reported Frequency of Symptoms*. Accessed: 2024-11-04. 2024. URL: https://data.cdc.gov/NCHS/Indicators-of-Anxiety-or-Depression-Based-on-Repor/8pt5-q6wp.

[134] L. Rubini, P. Soundarapandian und P. Eswaran. *Chronic Kidney Disease*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5G020. 2015.

[135] Benjamin Pineda und Jose De Jesus Flores Rivera. *Conversion predictors of Clinically Isolated Syndrome to Multiple Sclerosis in Mexican patients: a prospective study*. Mendeley Data, V1. 2023. DOI: 10.17632/8wk5hjx7x2.1.

[136] Shiva Borzooei und Aidin Tarokhian. *Differentiated Thyroid Cancer Recurrence*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5632J. 2023.

[137] Ugo Cesari u. a. *VOICED Dataset: A Database for Healthy and Pathological Voices*. University Hospital Federico II of Naples, Institute of High Performance Computing and Networking (ICAR-CNR). Available under the ODC Attribution License (ODC-By). 2017.

[138] Julia Kälsch u. a. „Normal liver enzymes are correlated with severity of metabolic syndrome in a large population based cohort“. In: *Scientific Reports* 5.1 (Aug. 2015), S. 13058. ISSN: 2045-2322. DOI: 10.1038/srep13058. URL: https://doi.org/10.1038/srep13058.

[139] Axel Schmermund u. a. „Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: Rationale and design of the Heinz Nixdorf RECALL Study“. In: *American Heart Journal* 144.2 (2002), S. 212–218. ISSN: 0002-8703. DOI: https://doi.org/10.1067/mhj.2002.123579. URL: https://www.sciencedirect.com/science/article/pii/S0002870302000698.

[140] M. Mostafizur Rahman und D. N. Davis. „Machine Learning-Based Missing Value Imputation Method for Clinical Datasets“. In: *IAENG Transactions on Engineering Technologies: Special Volume of the World Congress on Engineering 2012*. Hrsg. von Gi-Chul Yang, Sio-long Ao und Len Gelman. Dordrecht: Springer Netherlands, 2013, S. 245–257. ISBN: 978-94-007-6190-2. DOI: 10.1007/978-94-007-6190-2_19. URL: https://doi.org/10.1007/978-94-007-6190-2_19.

[141] Jorma Laurikkala u. a. „Informal identification of outliers in medical data“. In: *Fifth international workshop on intelligent data analysis in medicine and pharmacology*. Bd. 1. 1. 2000, S. 20–24.

[142] Robert J May, Holger R Maier und Graeme C Dandy. „Data splitting for artificial neural networks using SOM-based stratified sampling“. In: *Neural Networks* 23.2 (2010), S. 283–294.

[143] F. Pedregosa u. a. „Scikit-learn: Machine Learning in Python“. In: *Journal of Machine Learning Research* 12 (2011), S. 2825–2830.

[144] Adam Paszke u. a. „PyTorch: An Imperative Style, High-Performance Deep Learning Library“. In: *CoRR* abs/1912.01703 (2019). arXiv: 1912.01703. URL: http://arxiv.org/abs/1912.01703.

[145] Takuya Akiba u. a. „Optuna: A Next-generation Hyperparameter Optimization Framework“. In: *CoRR* abs/1907.10902 (2019). arXiv: 1907.10902. URL: http://arxiv.org/abs/1907.10902.

[146] Pankaj Chittora u. a. „Prediction of Chronic Kidney Disease - A Machine Learning Perspective“. In: *IEEE Access* 9 (2021), S. 17312–17334. DOI: 10.1109/ACCESS.2021.3053763.

[147] Judea Pearl. „The Do-Calculus Revisited“. In: *CoRR* abs/1210.4852 (2012). arXiv: 1210.4852. URL: http://arxiv.org/abs/1210.4852.

[148] Maria Nareklishvili, Nicholas Polson und Vadim Sokolov. *Generative Causal Inference*. 2023. arXiv: 2306.16096 [stat.ME]. URL: https://arxiv.org/abs/2306.16096.

A. Appendix

Tabelle A.1.: This table lists the modules used in the implementation of this thesis, including a description and version number for the respective module to ensure reproducibility.

Module	Description	Version
numpy	Python library for numerical computing, providing support for arrays, matrices, and mathematical functions.	1.24.3
pandas	Python library for data manipulation and analysis, providing data structures like DataFrames for handling structured data.	2.0.3
scikit-learn	Machine learning library for Python, providing simple and efficient tools for data analysis and predictive modeling.	1.3.0
scipy	Open-source Python library for scientific and technical computing, offering modules for optimization, integration, and statistics.	1.10.1
statsmodels	Python library for statistical modeling and hypothesis testing, including linear models, time series analysis, and statistical tests.	0.13.5
xgboost	Machine learning library for gradient boosting, optimized for performance and speed, commonly used for structured/tabular data.	2.0.3
matplotlib	Comprehensive library for creating static, animated, and interactive visualizations in Python.	3.6.3
seaborn	Python data visualization library based on matplotlib, providing a high-level interface for drawing statistical graphics.	0.13.2
cuda	NVIDIA's parallel computing architecture, enabling the use of GPUs for general-purpose computing tasks.	12.6
joblib	Python library for efficient serialization of objects and parallel computation.	1.4.2

Continued on next page

Tabelle A.1 – continued from previous page

Module	Description	Version
optuna	Hyperparameter optimization framework for machine learning, supporting optimization algorithms like TPE, CMA-ES, and more.	3.6.1
pip	Python package installer for installing and managing software packages written in Python.	24.0
pysoundfile	Python library for reading and writing sound files, supporting formats such as WAV, FLAC, and others.	0.12.1
python	The programming language used for implementation and analysis.	3.8.19
pytorch	Deep learning framework for building and training neural networks, with support for GPU acceleration.	2.4.0
pytorch-cuda	PyTorch extension providing GPU support for CUDA-enabled devices.	12.4
causal-learn	Python library for causal discovery and modeling, supporting algorithms like FCI and PC for causal inference.	0.1.3.8
pydot	Python interface to Graphviz, enabling the creation and manipulation of graphs and diagrams.	3.0.2
python-graphviz	Python interface to the Graphviz graph-drawing software, used for creating network diagrams and other graph-based visualizations.	0.20.3
torchaudio	PyTorch library for audio processing, including I/O functions and pre-built transforms for audio data.	2.4.0
torchvision	PyTorch library for computer vision tasks, providing datasets, models, and transforms for image processing.	0.19.0
librosa	Python library for audio and music analysis, providing tools for working with audio data, including feature extraction, time-frequency analysis, and audio I/O.	0.10.2

Tabelle A.2.: This table presents the optimal hyperparameters identified through the parameter search algorithm optuna for each dataset, specifically for the neural network models.

Disease	Model Type	Learning Rate	Batch Size	Epochs	Scheduler Factor	Scheduler Patience
Alzheimer	Both Models	0.000579	8	62	-	-
Breast Cancer	All Features	0.006445	8	71	0.451	19
	Non-Invasive	0.000307	4	95	0.329	7
Depression	Non-Invasive	0.000443	16	64	-	-
Heart Disease	All Features	0.000426	10	103	-	-
	Non-Invasive	0.002177	8	75	-	-
CKD	All Features	0.000189	8	91	-	-
	Non-Invasive	0.000141	32	46	-	-
Multiple Sclerosis	All Features	0.000697	10	122	-	-
	Non-Invasive	0.000972	4	119	-	-
Differentiated Thyroid Cancer Recurrence	All Features	0.0197	10	48	0.232	18
	Non-Invasive	0.00435	10	90	0.725	6

Tabelle A.3.: The following table presents the results of the feature importance analysis based on the random forest from the conducted model pipeline on the heart disease dataset.

Feature	**Importance**	**Invasive**
Slope of the peak exercise ST segment	0.3039	False
Exercise induced angina	0.1438	False
Oldpeak	0.1280	False
Chest pain type	0.1118	False
Serum cholesterol	0.0864	True
Maximum heart rate achieved	0.0762	False
Age	0.0516	False
Sex	0.0364	False
Resting blood pressure	0.0339	False
Fasting blood sugar	0.0177	True
Resting electrocardiographic results	0.0101	False

Tabelle A.4.: The following table presents the results of the feature importance analysis based on the random forest from the conducted model pipeline on the alzheimer disease dataset.

Feature	**Importance**	**Invasive**
Functional Assessment	0.2453	False
MMSE	0.2128	False
ADL	0.1964	False
Memory Complaints	0.1597	False
Behavioral Problems	0.1497	False
Cholesterol HDL	0.0077	True
Cholesterol Triglycerides	0.0045	True
Diet Quality	0.0042	False
Physical Activity	0.0037	False
Cholesterol LDL	0.0035	True
Age	0.0024	False
BMI	0.0020	False
Sleep Quality	0.0018	False
Cholesterol Total	0.0016	True
Alcohol Consumption	0.0013	False

Tabelle A.5.: The following table presents the results of the feature importance analysis based on the random forest from the conducted model pipeline on the breast cancer dataset.

Feature	**Importance**	**Invasive**
Tumor Size (cm)	0.3451	True
Inv-Nodes	0.2913	True
Metastasis	0.1719	True
Age	0.1093	False
Breast Quadrant	0.0337	False
Menopause	0.0278	False
History	0.0122	False
Breast	0.0086	False

Tabelle A.6.: The following table presents the results of the feature importance analysis based on the random forest from the conducted model pipeline on the depression dataset.

Feature	**Importance**	**Invasive**
Confidence Interval Midpoint	0.3946	False
Confidence Interval Lower	0.2567	False
Confidence Interval Upper	0.2333	False
Subgroup	0.0270	False
Phase	0.0261	False
Time Period	0.0203	False
Confidence Interval Range	0.0167	False
Group	0.0150	False
Indicator	0.0075	False
State	0.0024	False
Time Period Label	0.0004	False

Tabelle A.7.: The following table presents the results of the feature importance analysis based on the random forest from the conducted model pipeline on the chronic kidney disease dataset.

Feature	**Importance**	**Invasive**
Haemoglobin	0.1735	True
Specific Gravity	0.1733	False
Serum Creatinine	0.1559	True
Albumin	0.1096	True
Packed Cell Volume	0.0964	True
Hypertension	0.0709	False
Red Blood Cell Count	0.0478	True
Blood Glucose Random	0.0405	True
Diabetes Mellitus	0.0397	False
Blood Urea	0.0171	True
Appetite	0.0149	False
Red Blood Cells	0.0119	True

Tabelle A.8.: The following table presents the results of the feature importance analysis based on the random forest from the conducted model pipeline on the multiple sclerosis dataset.

Feature	**Importance**	**Invasive**
Periventricular MRI	0.3440	False
Initial Symptom	0.1876	False
Infratentorial MRI	0.1095	False
Schooling	0.0691	False
Oligoclonal Bands	0.0606	True
Age	0.0472	False
Gender	0.0381	False
Breastfeeding	0.0279	False
Varicella	0.0265	False
Cortical MRI	0.0243	False

Tabelle A.9.: The following table presents the results of the feature importance analysis based on the random forest from the conducted model pipeline on the pathological voice dataset.

Feature	**Importance**	**Invasive**
Voice Handicap Index (VHI) Score	0.1153	False
Carbonated Beverages	0.0594	False
MFCCS4	0.0389	False
Age	0.0323	False
Tomatoes	0.0247	False
MFCCS2	0.0246	False
Cups of Coffee Consumed Per Day	0.0241	False
Cigarettes Smoked Per Day	0.0240	False
MFCCS30	0.0233	False
MFCCS36	0.0224	False
MFCCS23	0.0224	False
MFCCS34	0.0220	False
MFCCS3	0.0214	False
MFCCS11	0.0211	False
MFCCS14	0.0207	False
MFCCS31	0.0201	False
Grams of Chocolate Consumed Per Day	0.0196	False
MFCCS9	0.0183	False
Smoker	0.0182	False
Citrus Fruits	0.0168	False

Continued on next page

Tabelle A.9 – continued from previous page

Feature	Importance	Invasive
MFCCS22	0.0168	False
MFCCS13	0.0159	False
Occupation Status	0.0158	False
MFCCS6	0.0155	False
MFCCS12	0.0150	False
MFCCS7	0.0148	False
MFCCS29	0.0140	False
Reflux Symptom Index (RSI) Score	0.0137	False
MFCCS38	0.0136	False
MFCCS27	0.0132	False
MFCCS16	0.0131	False
MFCCS40	0.0129	False
MFCCS18	0.0129	False
MFCCS21	0.0128	False
MFCCS39	0.0124	False
MFCCS25	0.0123	False
MFCCS8	0.0116	False
MFCCS26	0.0114	False
MFCCS5	0.0110	False
MFCCS24	0.0106	False
MFCCS41	0.0103	False
Gender	0.0098	False
Amount of Water Consumed Per Day	0.0093	False
MFCCS33	0.0093	False
MFCCS37	0.0092	False
MFCCS28	0.0091	False
MFCCS17	0.0090	False
MFCCS32	0.0088	False
MFCCS19	0.0082	False
MFCCS20	0.0077	False
MFCCS10	0.0075	False
MFCCS35	0.0072	False
MFCCS15	0.0071	False
Alcohol Consumption	0.0061	False
Amount of Glasses Consumed Per Day	0.0043	False
Chocolate	0.0041	False

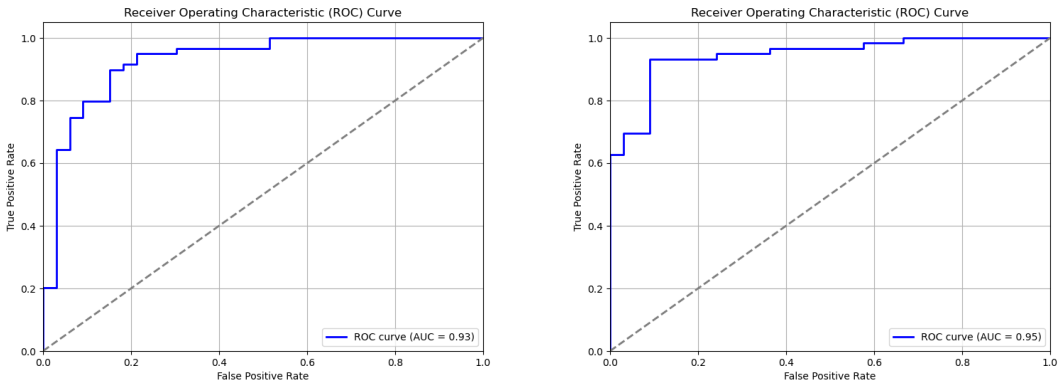
Continued on next page

Tabelle A.9 – continued from previous page

Feature	Importance	Invasive
Grams of Soft Cheese Consumed Per Day	0.0040	False
Number of Citrus Fruits Consumed Per Day	0.0030	False
Number of Glasses Containing Alcoholic Beverages Consumed	0.0028	False
Soft Cheese	0.0026	False
Coffee	0.0016	False

Tabelle A.10.: The following table presents the results of the feature importance analysis based on the random forest from the conducted model pipeline on the differentiated thyroid cancer recurrence dataset.

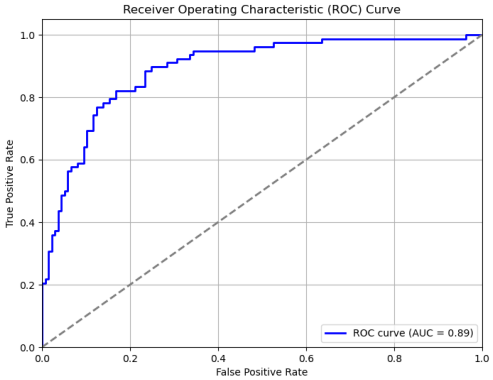
Feature	Importance	Invasive
Response to Treatment	0.4173	True
Risk	0.1974	False
N Stage	0.1153	True
Surgical Outcome	0.1027	True
Age	0.0491	False
BRAF Mutation	0.0182	True
Histology	0.0089	True
Gender	0.0057	False
Tumor Size	0.0046	True
Smoking History	0.0032	False



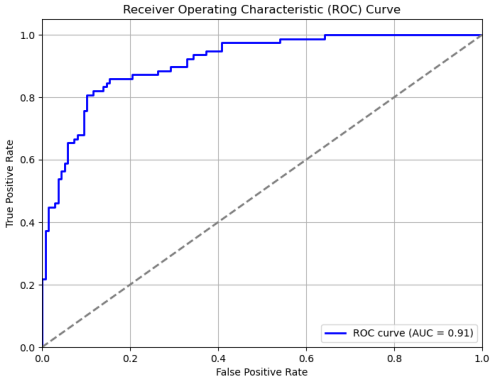
(a) Heart Disease ROC-Curve All Features

(b) Heart Disease ROC-Curve Non-Invasive

Abbildung A.1.: This figure compares the ROC curves for the all-features model (left) and the non-invasive model (right) on the heart disease dataset. The non-invasive model achieves a slightly higher area under the curve, indicating marginally better discriminatory performance.

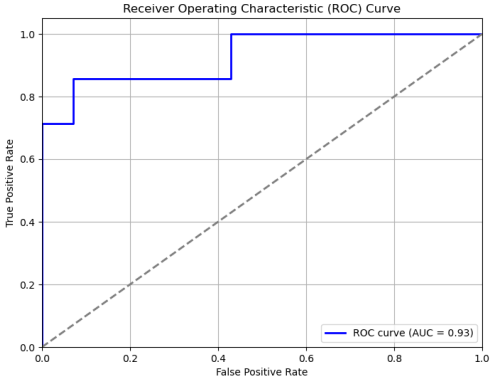


(a) Alzheimer ROC-Curve All Features

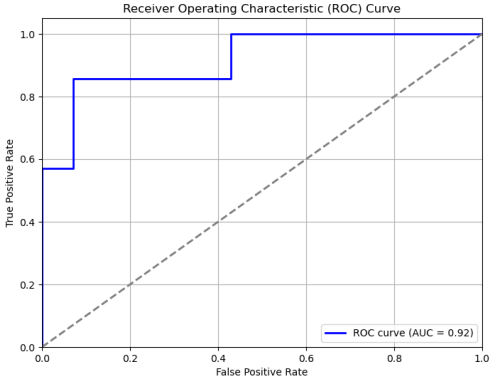


(b) Alzheimer ROC-Curve Non-Invasive

Abbildung A.2.: This figure compares the ROC curves for the all-features model (left) and the non-invasive model (right) on the alzheimer disease dataset. The non-invasive model achieves a slightly higher area under the curve, indicating marginally better discriminatory performance.

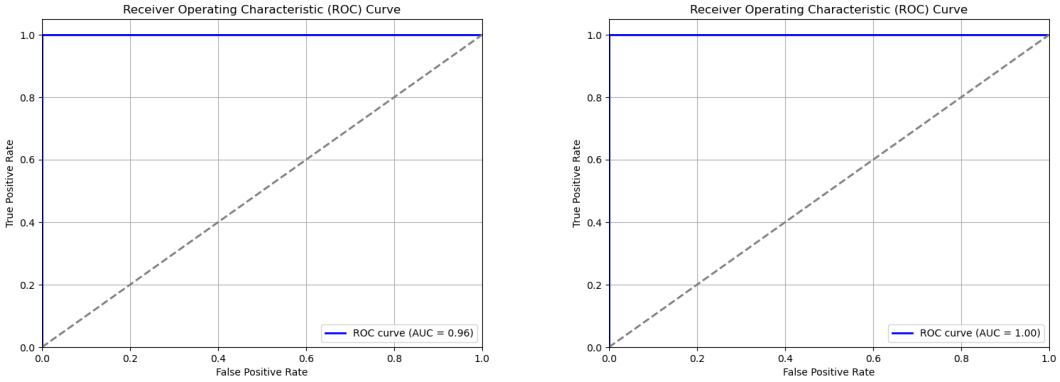


(a) Breast Cancer ROC-Curve All Features



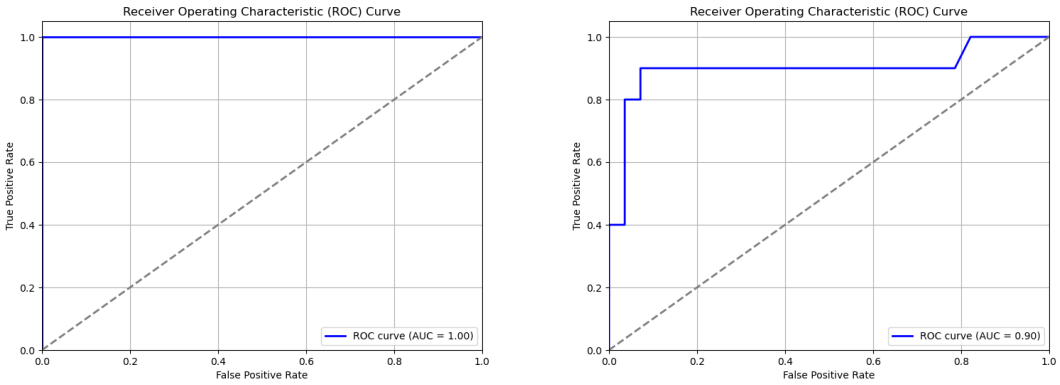
(b) Breast Cancer ROC-Curve Non-Invasive

Abbildung A.3.: This figure compares the ROC curves for the all-features model (left) and the non-invasive model (right) on the breast cancer dataset. The non-invasive model achieves a slightly lower area under the curve, indicating marginally worse discriminatory performance.



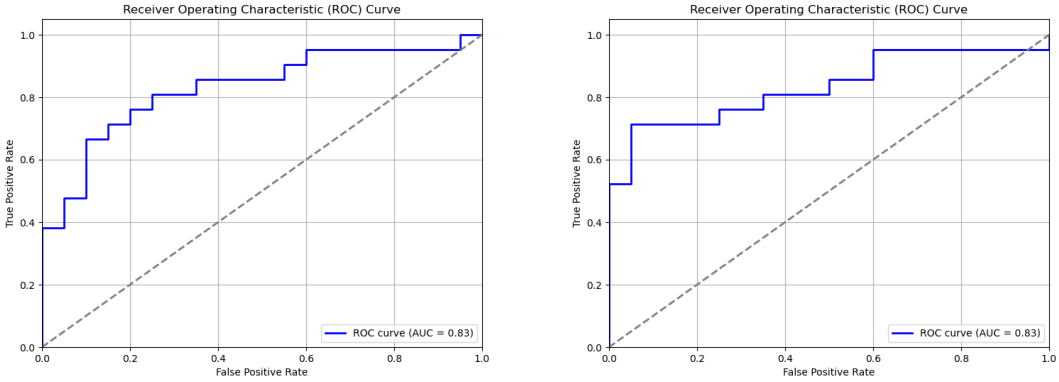
(a) Chronic Kidney Disease ROC-Curve All Features (b) Chronic Kidney Disease ROC-Curve Non-Invasive

Abbildung A.4.: This figure compares the ROC curves for the all-features model (left) and the non-invasive model (right) on the chronic kidney disease dataset. The non-invasive model achieves a slightly higher area under the curve, indicating marginally better discriminatory performance.



(a) Differentiated Thyroid Cancer Recurrence ROC-Curve All Features (b) Differentiated Thyroid Cancer Recurrence ROC-Curve Non-Invasive

Abbildung A.5.: This figure compares the ROC curves for the all-features model (left) and the non-invasive model (right) on the differentiated thyroid recurrence dataset. The non-invasive model achieves a lower area under the curve, indicating worse discriminatory performance.



(a) Multiple Sclerosis ROC-Curve All Features (b) Multiple Sclerosis ROC-Curve Non-Invasive

Abbildung A.6.: This figure compares the ROC curves for the all-features model (left) and the non-invasive model (right) on the multiple sclerosis dataset. Both models have almost the same are under the curve, indicating almost equal discriminatory performance.

Tabelle A.11.: This table lists the performance metrics for the non-invasive breast cancer model on the test set. The Neural Net has the best performance with almost consistent higher metrics.

Model	**ROC-AUC**	**Sensitivity**	**Specificity**	**F1-Score**	**Accuracy**
NeuralNet	0.918	0.857	0.857	0.799	0.857
Random Forest	0.867	0.857	0.785	0.750	0.809
XGBoost	0.867	0.857	0.785	0.750	0.809
KNN	0.928	0.714	0.928	0.769	0.857
GB Classifier	0.836	0.714	0.714	0.625	0.714
SVM	0.918	0.714	0.857	0.714	0.809

Tabelle A.12.: This table lists the performance metrics for the all features breast cancer model on the test set. The Neural Net and XGBoost have the most balanced performance compared to the other models.

Model	**ROC-AUC**	**Sensitivity**	**Specificity**	**F1-Score**	**Accuracy**
NeuralNet	0.928	0.857	0.928	0.857	0.904
Random Forest	0.918	0.714	1.000	0.833	0.904
XGBoost	0.918	0.714	1.000	0.833	0.904
KNN	0.918	0.714	1.000	0.833	0.904
GB Classifier	0.928	0.857	0.928	0.857	0.904
SVM	0.836	0.714	0.928	0.769	0.857

Tabelle A.13.: This table lists the performance metrics for the non-invasive heart disease model on the test set. The Neural Net has the best performance with consistent higher metrics.

Model	**ROC-AUC**	**Sensitivity**	**Specificity**	**F1-Score**	**Accuracy**
NeuralNet	0.945	0.915	0.909	0.931	0.912
Random Forest	0.888	0.898	0.878	0.913	0.891
XGBoost	0.856	0.864	0.848	0.887	0.858
KNN	0.839	0.830	0.848	0.867	0.837
GB Classifier	0.839	0.830	0.848	0.867	0.837
SVM	0.820	0.762	0.878	0.833	0.804

Tabelle A.14.: This table lists the performance metrics for the all features heart disease model on the test set. The Neural Net has the best performance with almost consistent higher metrics.

Model	**ROC-AUC**	**Sensitivity**	**Specificity**	**F1-Score**	**Accuracy**
NeuralNet	0.927	0.898	0.818	0.898	0.869
Random Forest	0.895	0.881	0.909	0.912	0.891
XGBoost	0.903	0.898	0.909	0.921	0.902
KNN	0.848	0.847	0.848	0.877	0.847
GB Classifier	0.895	0.881	0.909	0.912	0.891
SVM	0.837	0.796	0.878	0.854	0.826

Tabelle A.15.: This table lists the performance metrics for the non-invasive chronic kidney disease model on the test set. All models had perfect performance on the test set.

Model	**ROC-AUC**	**Sensitivity**	**Specificity**	**F1-Score**	**Accuracy**
NeuralNet	1.000	1.000	1.000	1.000	1.000
Random Forest	1.000	1.000	1.000	1.000	1.000
XGBoost	1.000	1.000	1.000	1.000	1.000
KNN	1.000	1.000	1.000	1.000	1.000
GB Classifier	1.000	1.000	1.000	1.000	1.000
SVM	1.000	1.000	1.000	1.000	1.000

Tabelle A.16.: This table lists the performance metrics for the all features chronic kidney disease model on the test set. The Random Forest, KNN and SVM models hat the best performance with perfect metrics on the test data.

Model	**ROC-AUC**	**Sensitivity**	**Specificity**	**F1-Score**	**Accuracy**
NeuralNet	0.964	0.928	1.000	0.962	0.950
Random Forest	1.000	1.000	1.000	1.000	1.000
XGBoost	0.9821	0.9643	1.000	0.9818	0.975
KNN	1.000	1.000	1.000	1.000	1.000
GB Classifier	0.9821	0.9643	1.000	0.9818	0.975
SVM	1.000	1.000	1.000	1.000	1.000

Tabelle A.17.: This table lists the performance metrics for the non-invasive multiple sclerosis model on the test set. The Neural Net, Gradient Boosting and SVM had the best performance metrics on the test set.

Model	**ROC-AUC**	**Sensitivity**	**Specificity**	**F1-Score**	**Accuracy**
NeuralNet	0.833	0.761	0.750	0.761	0.756
Random Forest	0.757	0.714	0.800	0.750	0.756
XGBoost	0.757	0.714	0.800	0.750	0.756
KNN	0.731	0.761	0.700	0.744	0.731
GB Classifier	0.833	0.761	0.750	0.761	0.756
SVM	0.833	0.761	0.750	0.761	0.756

Tabelle A.18.: This table lists the performance metrics for the all features multiple sclerosis model on the test set. The Neural Net and Gradient Boosting had the best performance metrics on the test set.

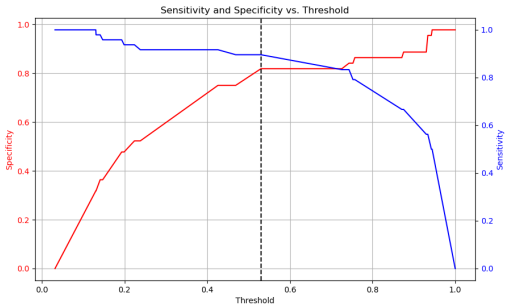
Model	**ROC-AUC**	**Sensitivity**	**Specificity**	**F1-Score**	**Accuracy**
NeuralNet	0.830	0.761	0.750	0.761	0.756
Random Forest	0.757	0.714	0.800	0.750	0.756
XGBoost	0.756	0.714	0.800	0.750	0.756
KNN	0.731	0.809	0.650	0.755	0.731
GB Classifier	0.830	0.761	0.750	0.761	0.756
SVM	0.731	0.761	0.700	0.744	0.731

Tabelle A.19.: This table lists the performance metrics for the non-invasive differentiated thyroid cancer recurrence model on the test set. The Neural Net had the best performance metrics on the test set.

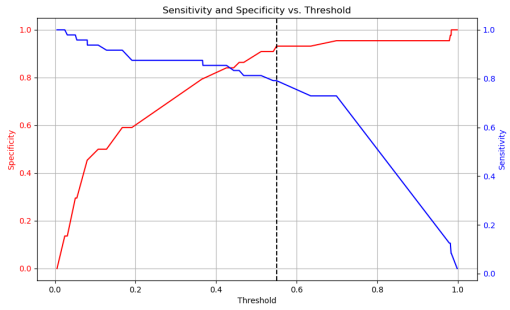
Model	**ROC-AUC**	**Sensitivity**	**Specificity**	**F1-Score**	**Accuracy**
NeuralNet	0.898	0.900	0.928	0.857	0.921
Random Forest	0.846	0.800	0.892	0.761	0.868
XGBoost	0.864	0.800	0.928	0.800	0.894
KNN	0.832	0.700	0.964	0.777	0.894
GB Classifier	0.864	0.800	0.928	0.800	0.894
SVM	0.750	0.500	1.000	0.666	0.868

Tabelle A.20.: This table lists the performance metrics for the all features differentiated thyroid cancer recurrence model on the test set. Every model except the KNN and Gradient boosting models had perfect metrics on the test set.

Model	**ROC-AUC**	**Sensitivity**	**Specificity**	**F1-Score**	**Accuracy**
NeuralNet	1.000	1.000	1.000	1.000	1.000
Random Forest	1.000	1.000	1.000	1.000	1.000
XGBoost	1.000	1.000	1.000	1.000	1.000
KNN	0.942	0.700	1.000	0.823	0.921
GB Classifier	1.000	0.900	1.000	0.947	0.973
SVM	1.000	1.000	1.000	1.000	1.000

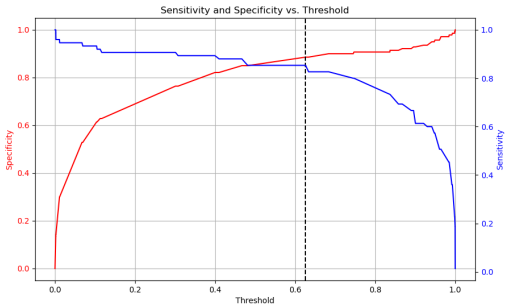


(a) Heart Disease Cutoff All Features

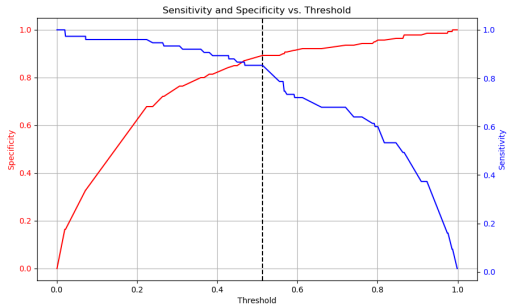


(b) Heart Disease Cutoff Non-Invasive

Abbildung A.7.: This figure shows the optimal threshold determined by balancing sensitivity and specificity for the heart disease dataset.

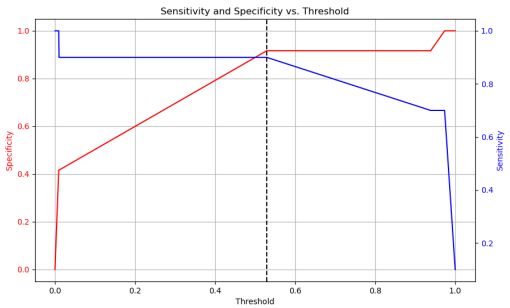


(a) Alzheimer Cutoff All Features

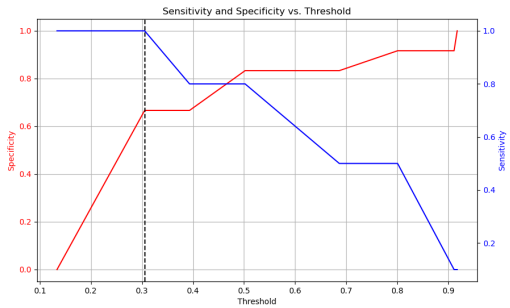


(b) Alzheimer Cutoff Non-Invasive

Abbildung A.8.: This figure shows the optimal threshold determined by balancing sensitivity and specificity for the alzheimer disease dataset.

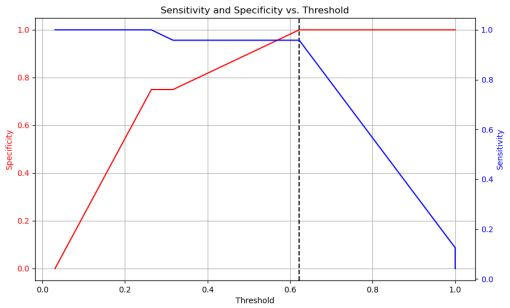


(a) Breast Cancer Cutoff All Features

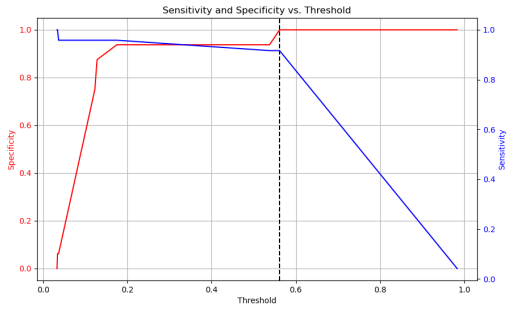


(b) Breast Cancer Cutoff Non-Invasive

Abbildung A.9.: This figure shows the optimal threshold determined by balancing sensitivity and specificity for the breast cancer dataset.

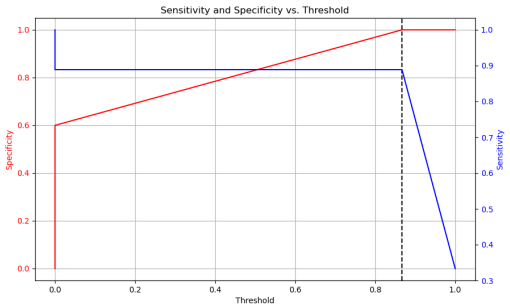


(a) Chronic Kidney Disease Cutoff All Features

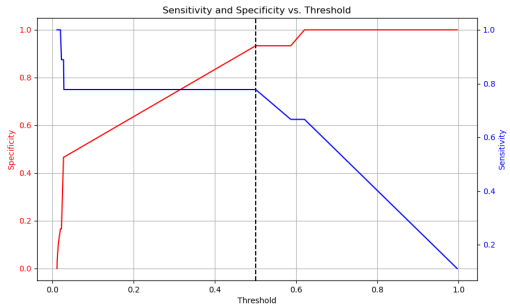


(b) Chronic Kidney Disease Cutoff Non-Invasive

Abbildung A.10.: This figure shows the optimal threshold determined by balancing sensitivity and specificity for the chronic kidney disease dataset.

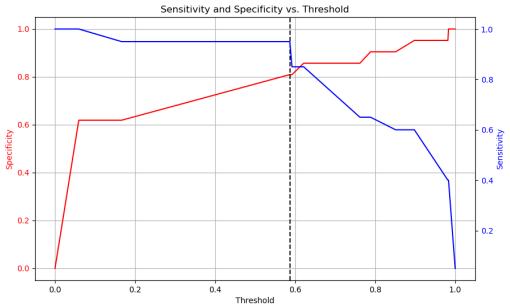


(a) Differentiated Thyroid Cancer Recurrence Cutoff All Features

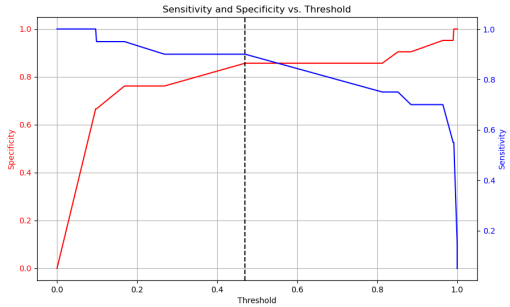


(b) Differentiated Thyroid Cancer Recurrence Cutoff Non-Invasive

Abbildung A.11.: This figure shows the optimal threshold determined by balancing sensitivity and specificity for the differentiated thyroid cancer recurrence dataset.



(a) Multiple Sclerosis Cutoff All Features



(b) Multiple Sclerosis Cutoff Non-Invasive

Abbildung A.12.: This figure shows the optimal threshold determined by balancing sensitivity and specificity for the multiple sclerosis dataset.

Tabelle A.21.: This table presents the score Calibration Errors for Each Dataset after fitting a probability distribution to the class predictions.

Dataset	Model	Class	ECE	MCE	LCE
Alzheimer	Non-Invasive	0	0.4272	0.9729	0.2124
		1	0.3572	0.7768	0.2200
	All Features	0	0.5690	0.9901	0.1342
		1	0.2307	0.8316	0.0863
Breast Cancer	Non-Invasive	0	0.3578	0.7869	0.3533
		1	0.3909	0.8920	0.3858
	All Features	0	0.3329	0.8494	0.0646
		1	0.1642	0.8779	0.0798
Heart Disease	Non-Invasive	0	0.5662	0.9567	0.2196
		1	0.4649	0.7337	0.1753
	All Features	0	0.5834	0.9359	0.2998
		1	0.3949	0.8200	0.1222
Chronic Kidney	Non-Invasive	0	0.0930	0.4714	0.0869
		1	0.4125	0.8107	0.1400
	All Features	0	0.5464	0.9748	0.1999
		1	0.3031	0.8673	0.0855
Multiple Sclerosis	Non-Invasive	0	0.3240	0.8479	0.0253
		1	0.2423	0.8354	0.0771
	All Features	0	0.2333	0.8981	0.1394
		1	0.3024	0.8392	0.1180
Thyroid Recurrence	Non-Invasive	0	0.4710	0.8441	0.1166
		1	0.2366	0.8698	0.2851
	All Features	0	0.6718	0.9706	0.0054
		1	0.0167	0.8162	0.0081

Tabelle A.22.: This table presents the calibration Errors for Each Dataset after fitting a probability distribution to the class predictions.

Dataset	Model	Class	ECE	MCE	LCE
Alzheimer	Non-Invasive	0	0.3992	0.8534	0.8314
		1	0.3141	0.8861	0.1151
	All Features	0	0.1643	0.9234	0.8951
		1	0.0794	0.9246	0.0417
Breast Cancer	Non-Invasive	0	0.6047	0.8159	0.6878
		1	0.3722	0.7275	0.3132
	All Features	0	0.0732	0.8440	0.9595
		1	0.0535	0.5442	0.0449
Heart Disease	Non-Invasive	0	0.4297	0.8489	0.8526
		1	0.3797	0.8538	0.1307
	All Features	0	0.0047	0.0060	0.9940
		1	0.5405	0.5436	0.5457
Chronic Kidney	Non-Invasive	0	0.1237	0.8872	0.9629
		1	0.0665	0.9277	0.0203
	All Features	0	0.0163	0.0164	0.9836
		1	0.3533	0.3535	0.3538
Multiple Sclerosis	Non-Invasive	0	0.0116	0.0119	0.9878
		1	0.0093	0.0093	0.0093
	All Features	0	0.0906	0.9368	0.9532
		1	0.2593	0.9047	0.1320
Thyroid Recurrence	Non-Invasive	0	0.1646	0.9050	0.8566
		1	0.1174	0.6224	0.1077
	All Features	0	0.0092	0.0092	0.9908
		1	0.0	0.0	0.0

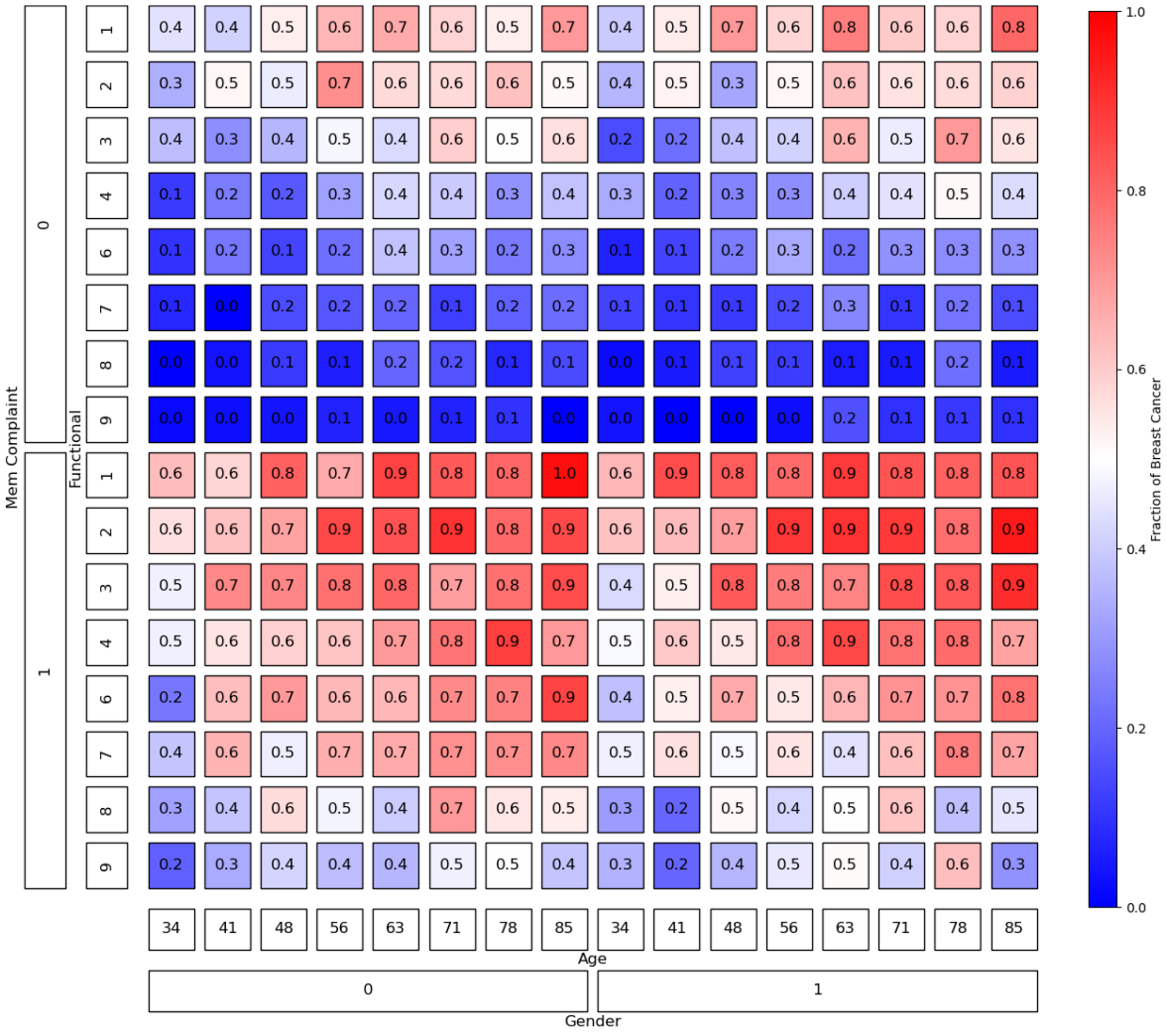


Abbildung A.13.: This figure illustrates the patterns learned by the non-invasive neural net from the Alzheimer disease dataset based on the predictions. The depth of the colors represent the fraction of Alzheimer.

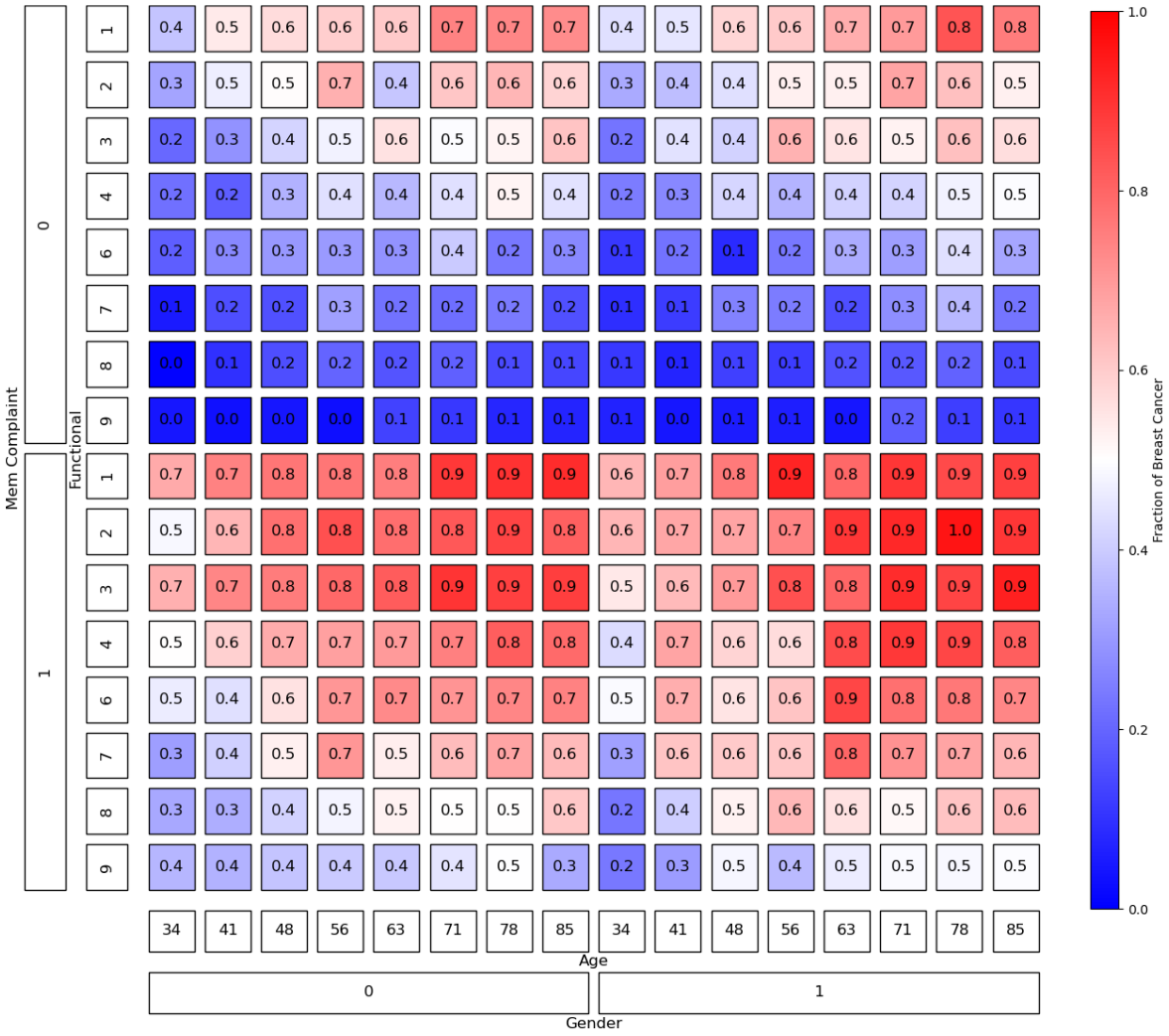


Abbildung A.14.: This figure illustrates the patterns learned by the non-invasive neural net from the Alzheimer disease dataset based on the calibrations. The depth of the colors represent the fraction of Alzheimer.

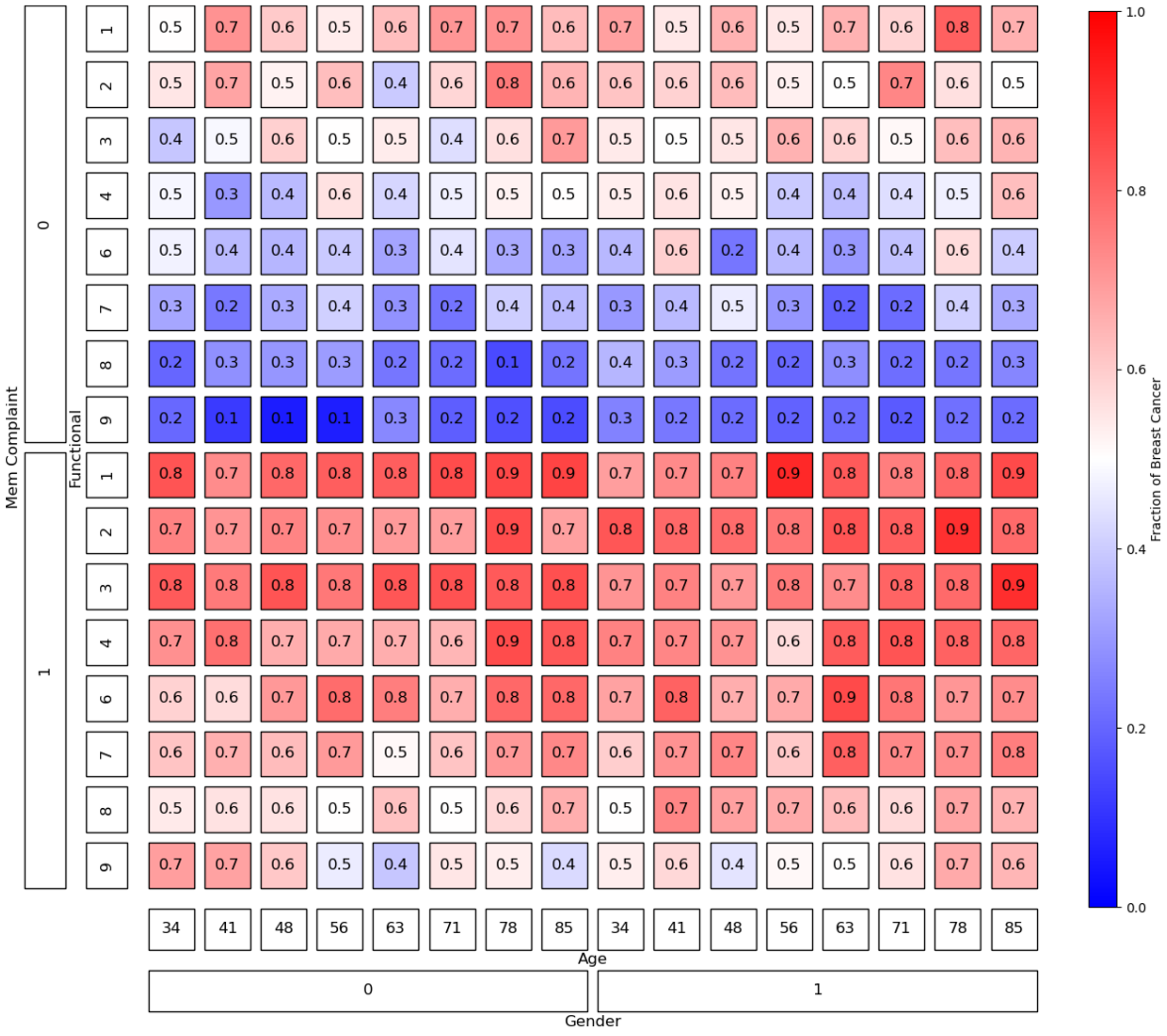


Abbildung A.15.: This figure illustrates the patterns learned by the all features neural net from the Alzheimer disease dataset based on the predictions. The depth of the colors represent the fraction of Alzheimer.

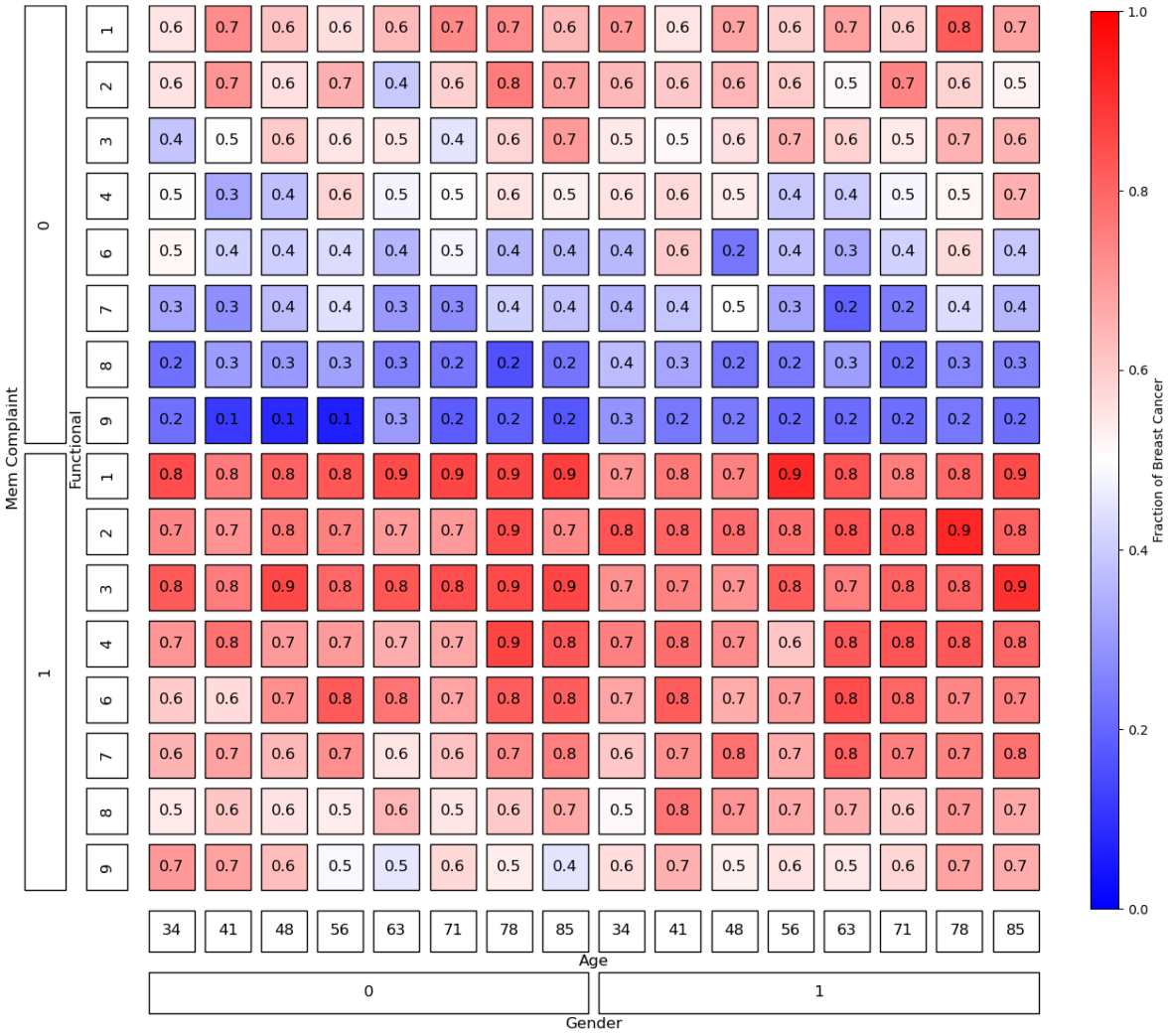


Abbildung A.16.: This figure illustrates the patterns learned by the all features neural net from the Alzheimer disease dataset based on the calibrations. The depth of the colors represent the fraction of Alzheimer.

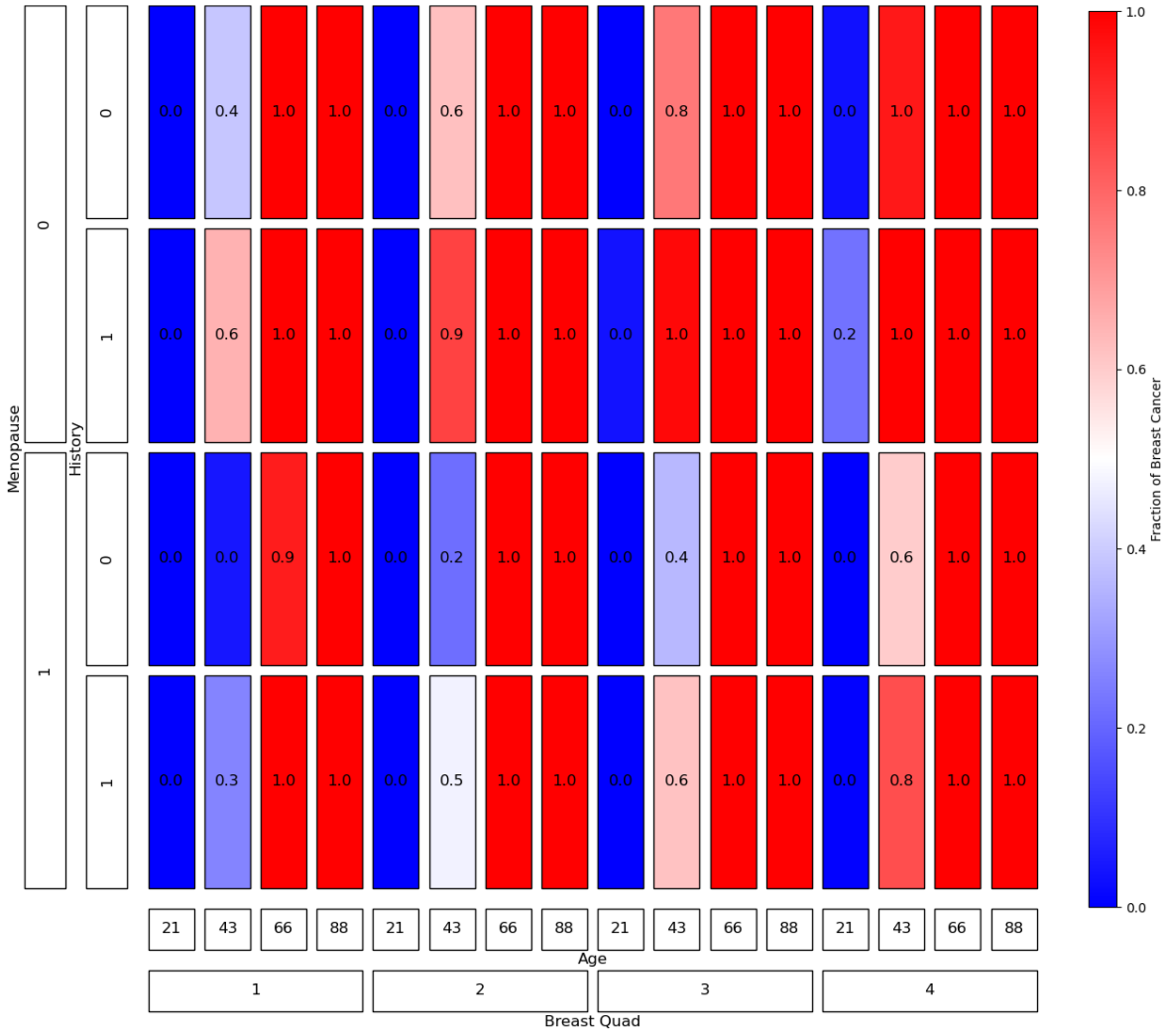


Abbildung A.17.: This figure illustrates the patterns learned by the non-invasive neural net from the Breast Cancer dataset based on the predictions. The depth of the colors represent the fraction of Breast Cancer.

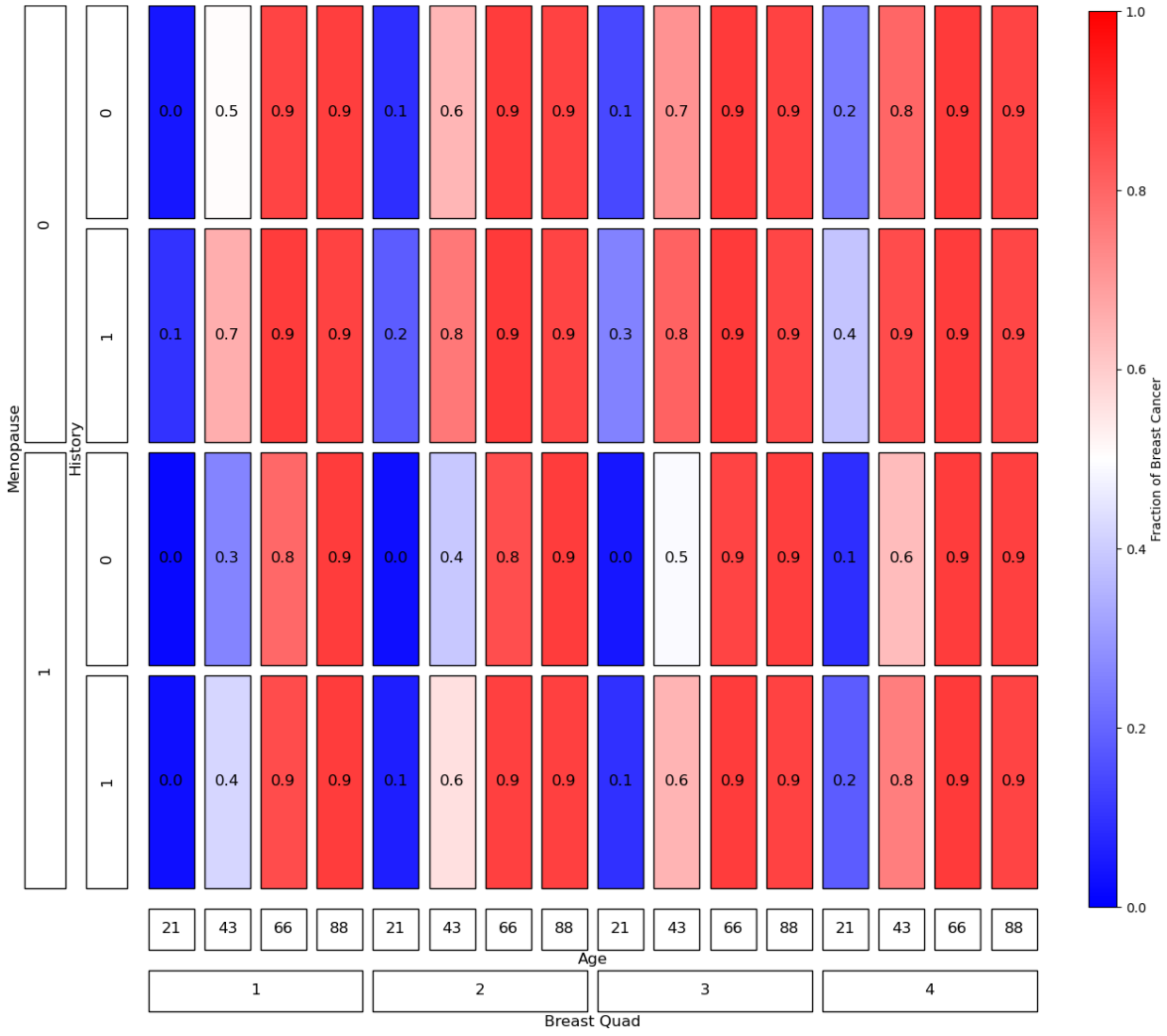


Abbildung A.18.: This figure illustrates the patterns learned by the non-invasive neural net from the Breast Cancer dataset based on the calibrations. The depth of the colors represent the fraction of Breast Cancer.

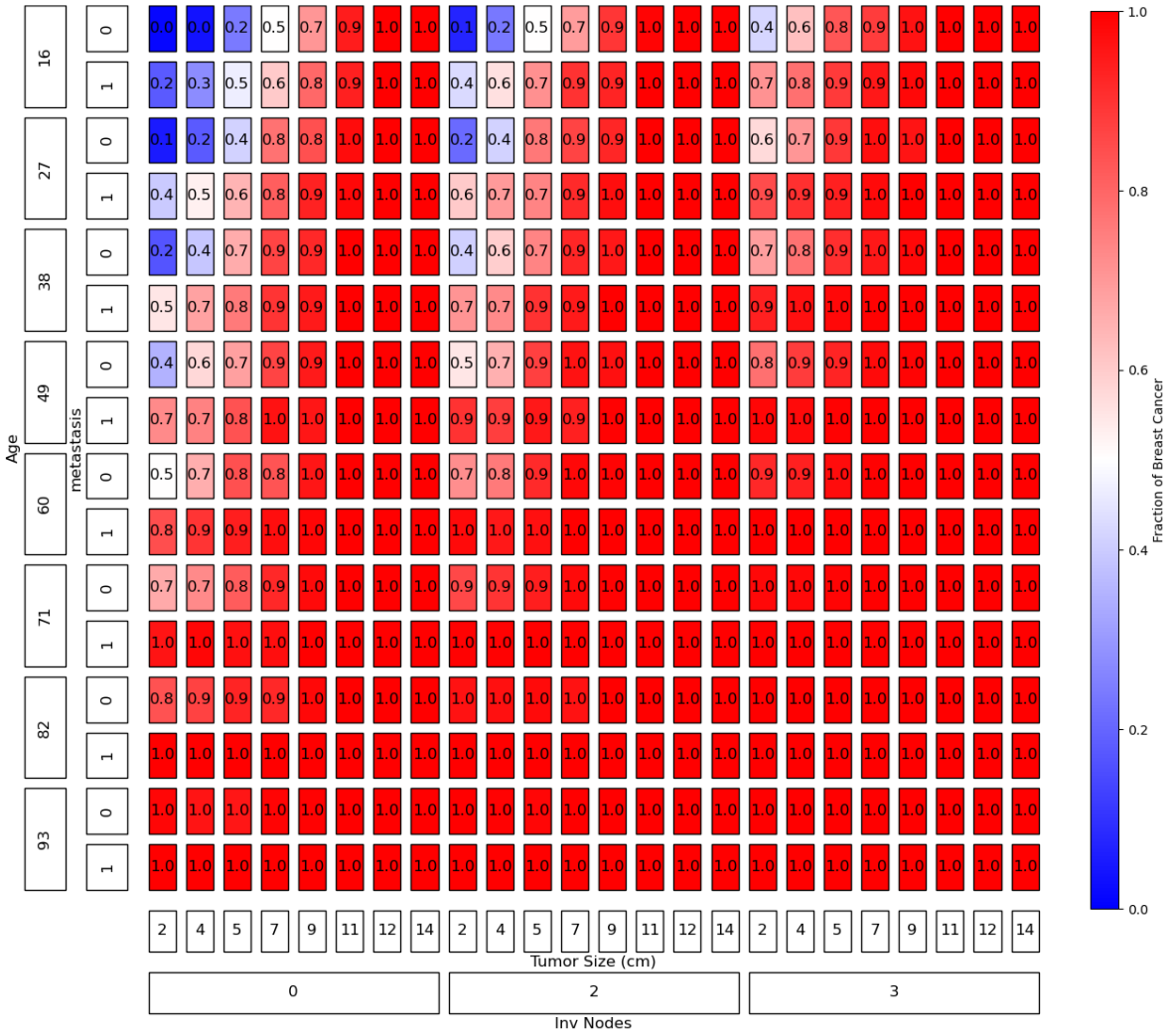


Abbildung A.19.: This figure illustrates the patterns learned by the all features neural net from the Breast Cancer dataset based on the predictions. The depth of the colors represent the fraction of Breast Cancer.

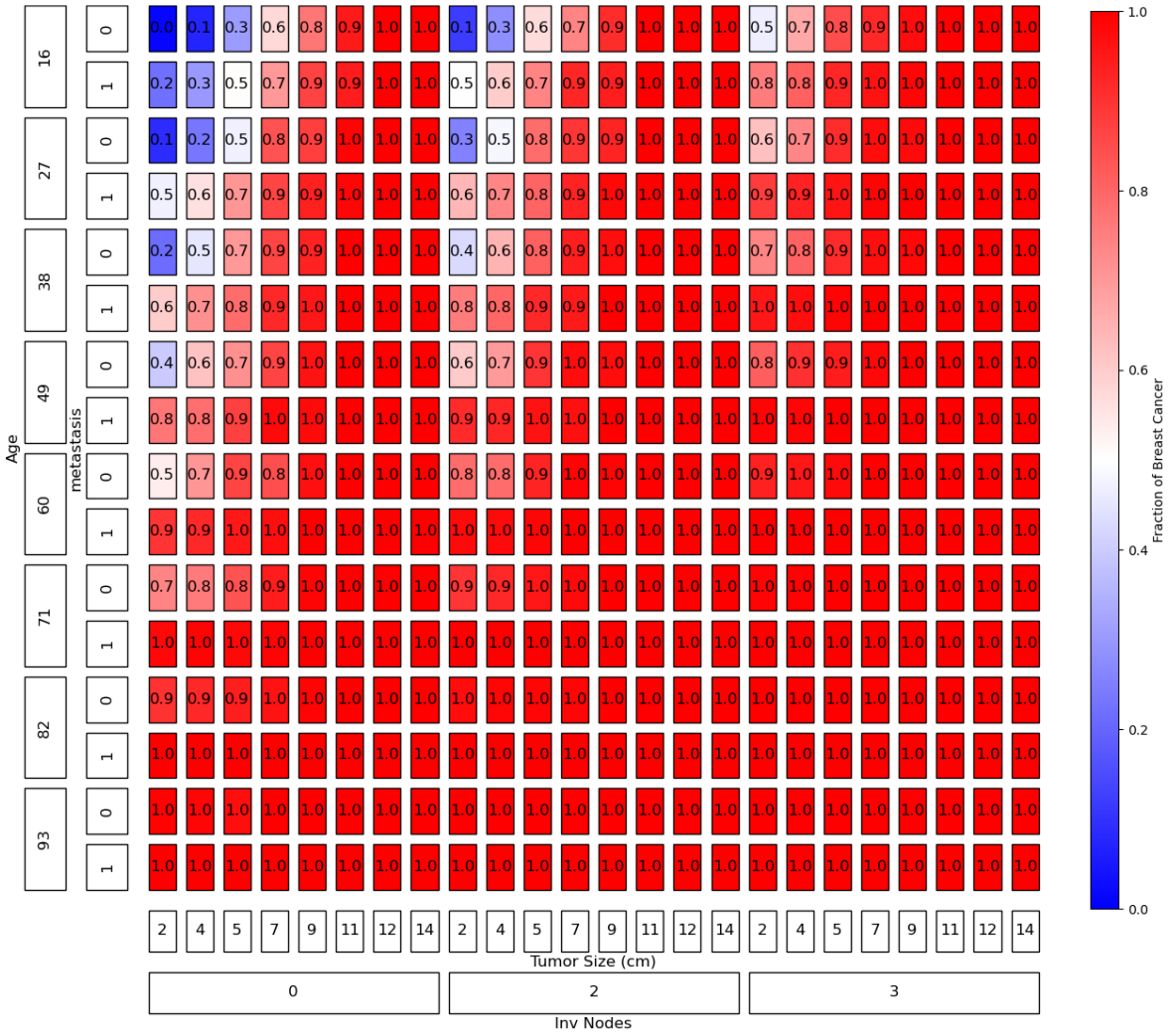


Abbildung A.20.: This figure illustrates the patterns learned by the all features neural net from the Breast Cancer dataset based on the calibrations. The depth of the colors represent the fraction of Breast Cancer.

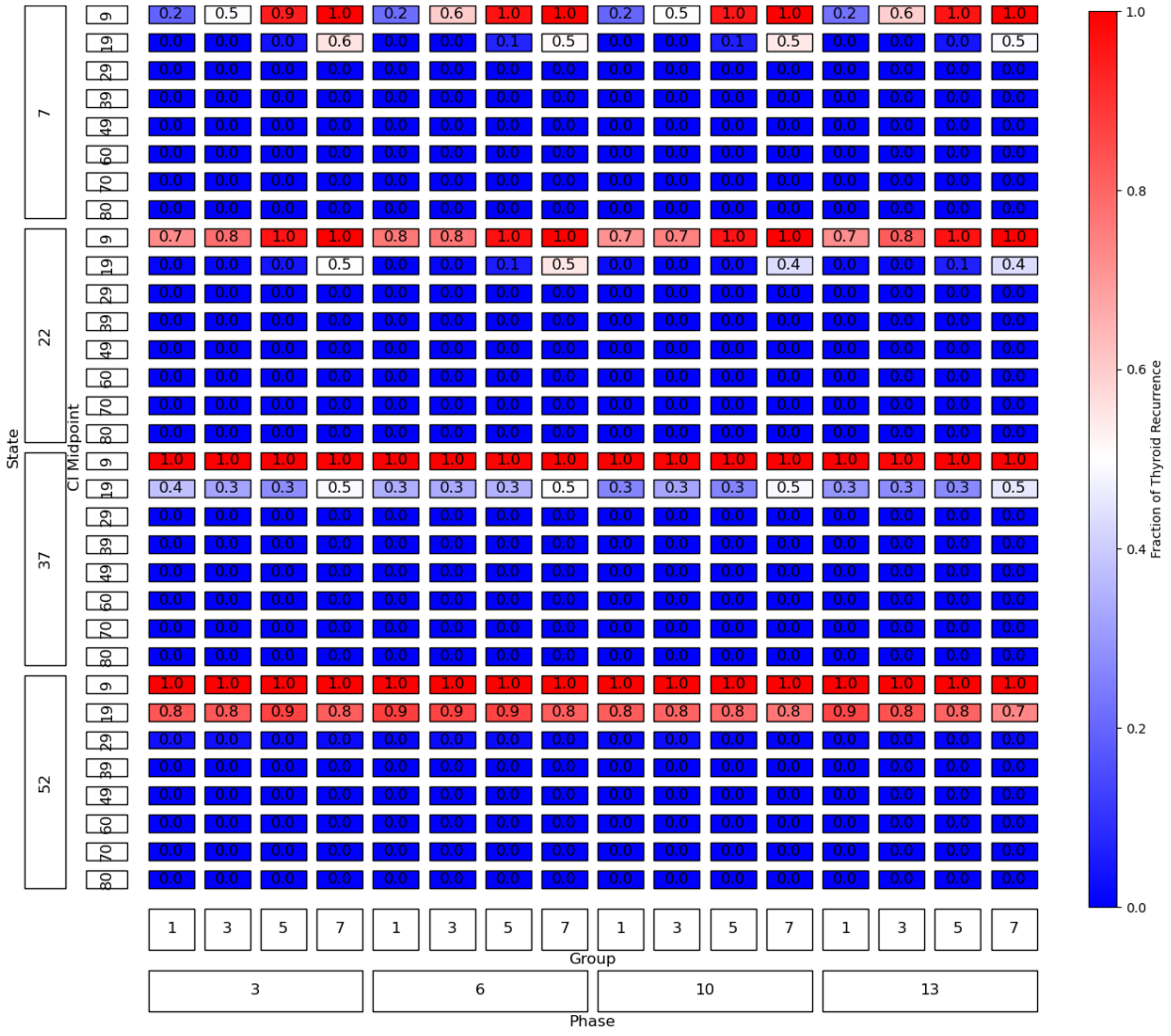


Abbildung A.21.: This figure illustrates the patterns learned by the neural net from the depression dataset based on the predictions for class 0. The depth of the colors represent the fraction of Depression.

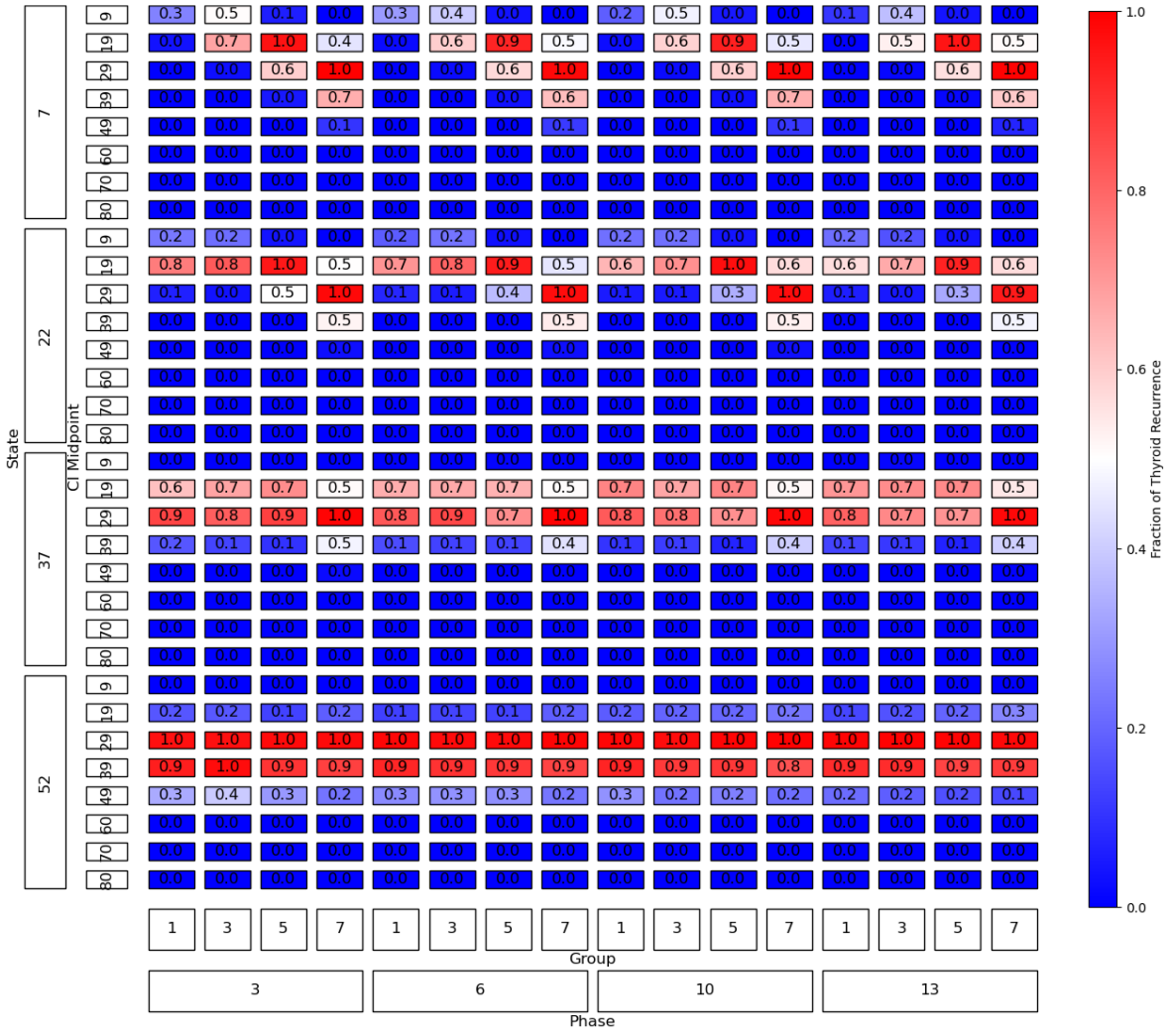


Abbildung A.22.: This figure illustrates the patterns learned by the neural net from the depression dataset based on the predictions for class 1. The depth of the colors represent the fraction of Depression.

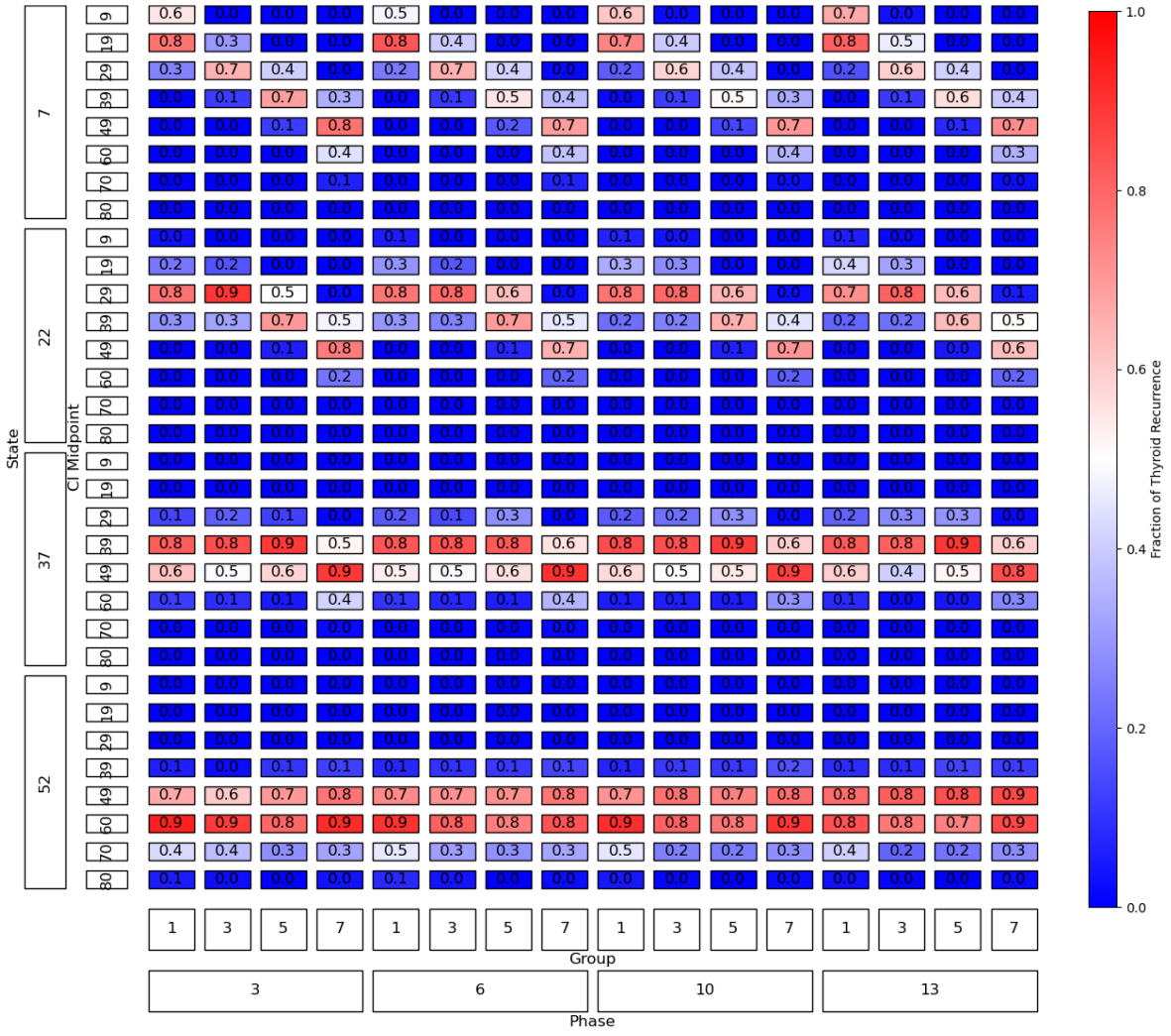


Abbildung A.23.: This figure illustrates the patterns learned by the neural net from the depression dataset based on the predictions for class 2. The depth of the colors represent the fraction of Depression.

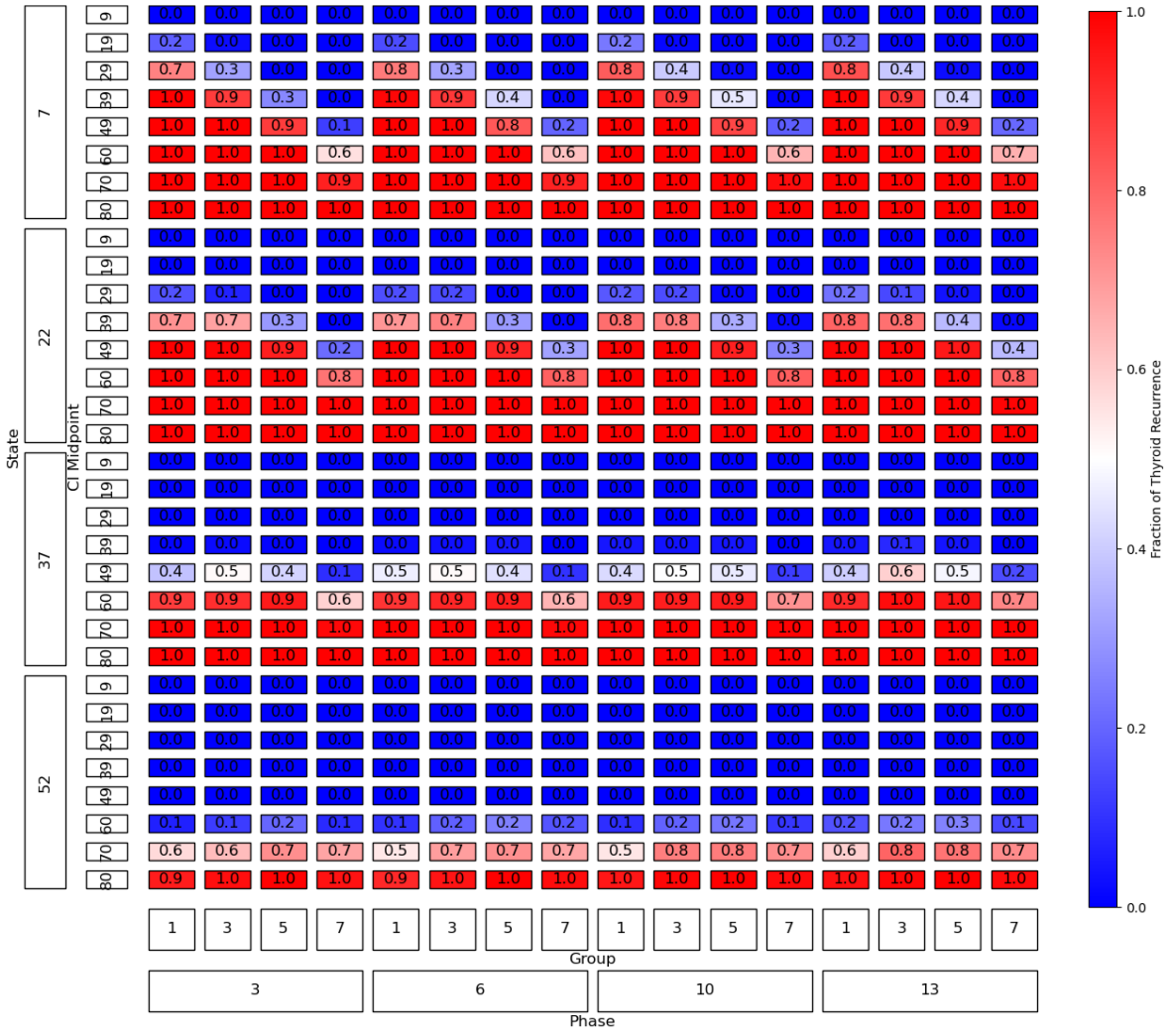


Abbildung A.24.: This figure illustrates the patterns learned by the neural net from the depression dataset based on the predictions for class 3. The depth of the colors represent the fraction of Depression.

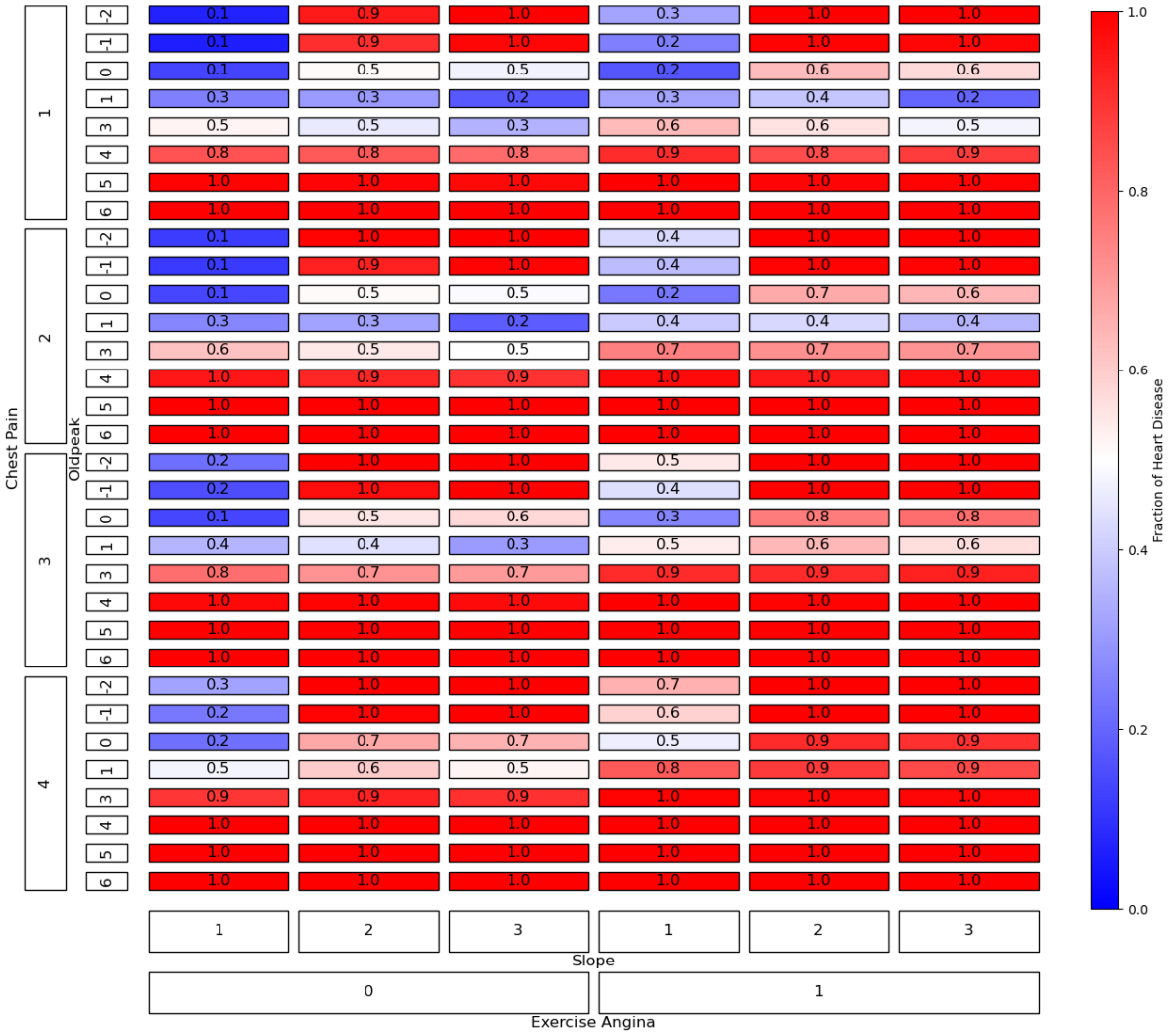


Abbildung A.25.: This figure illustrates the patterns learned by the non-invasive neural net from the heart disease dataset based on the predictions. The depth of the colors represent the fraction of heart disease.

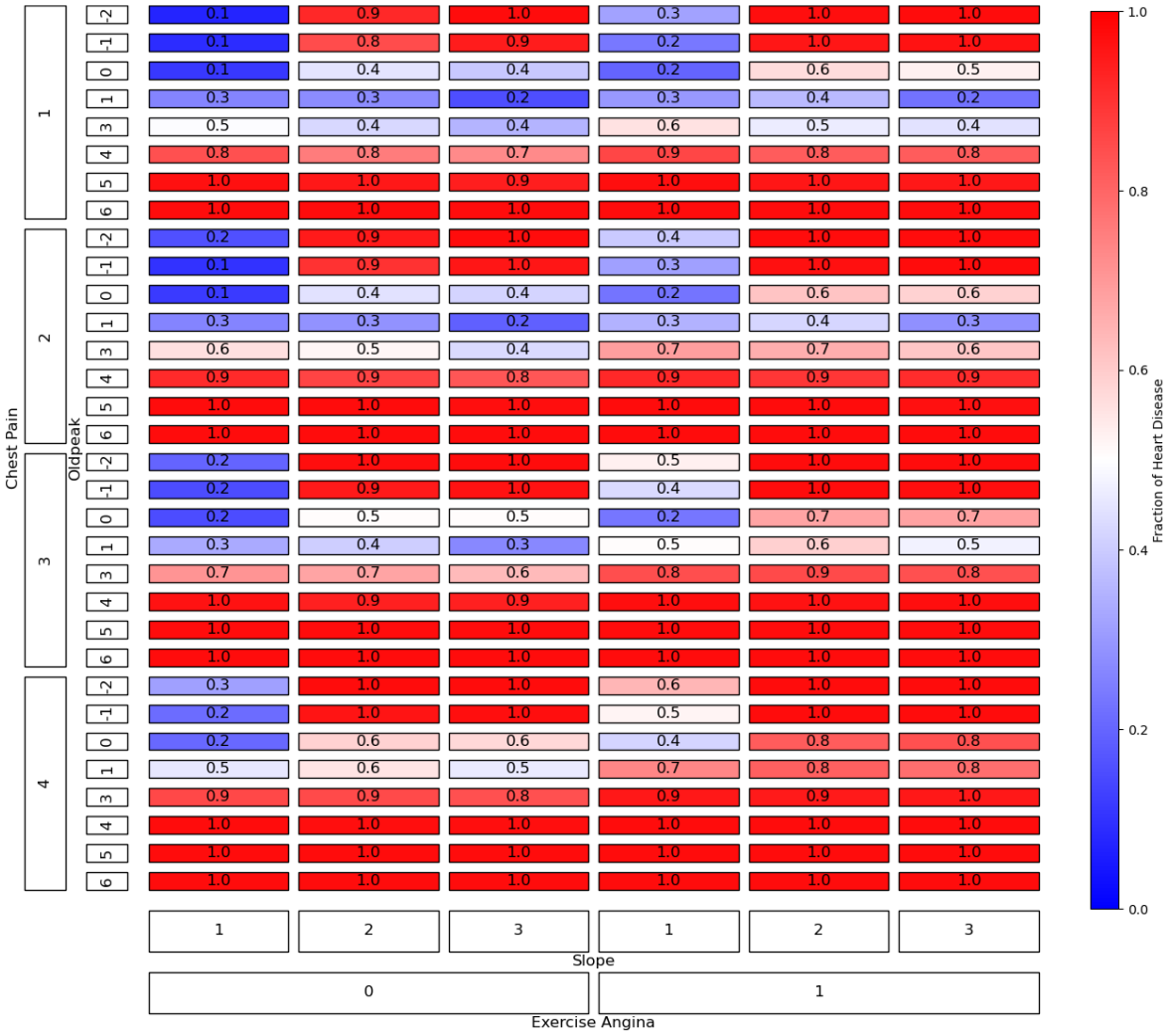


Abbildung A.26.: This figure illustrates the patterns learned by the non-invasive neural net from the heart disease dataset based on the calibrations. The depth of the colors represent the fraction of heart disease.

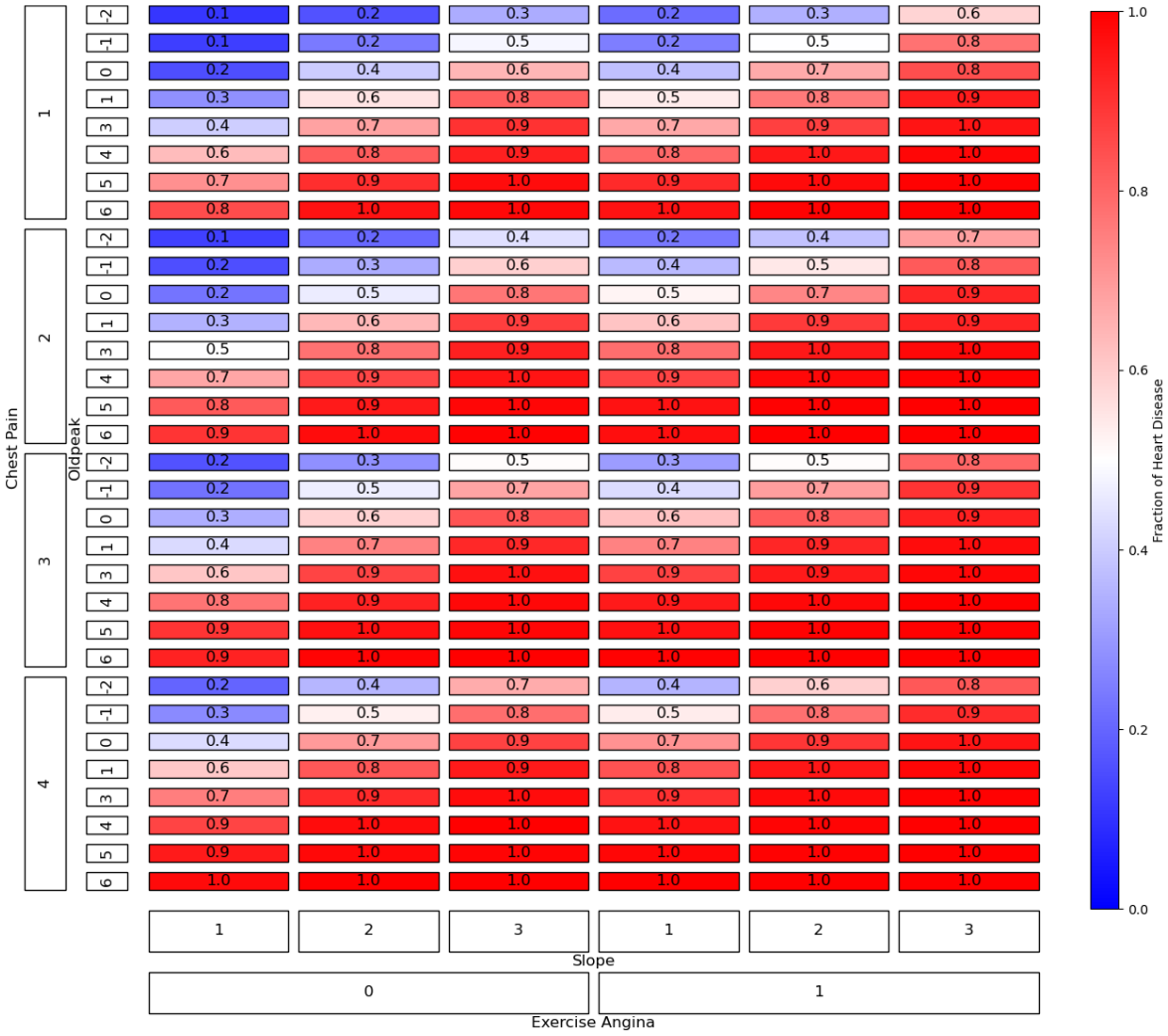


Abbildung A.27.: This figure illustrates the patterns learned by the all features neural net from the heart disease dataset based on the predictions. The depth of the colors represent the fraction of heart disease.

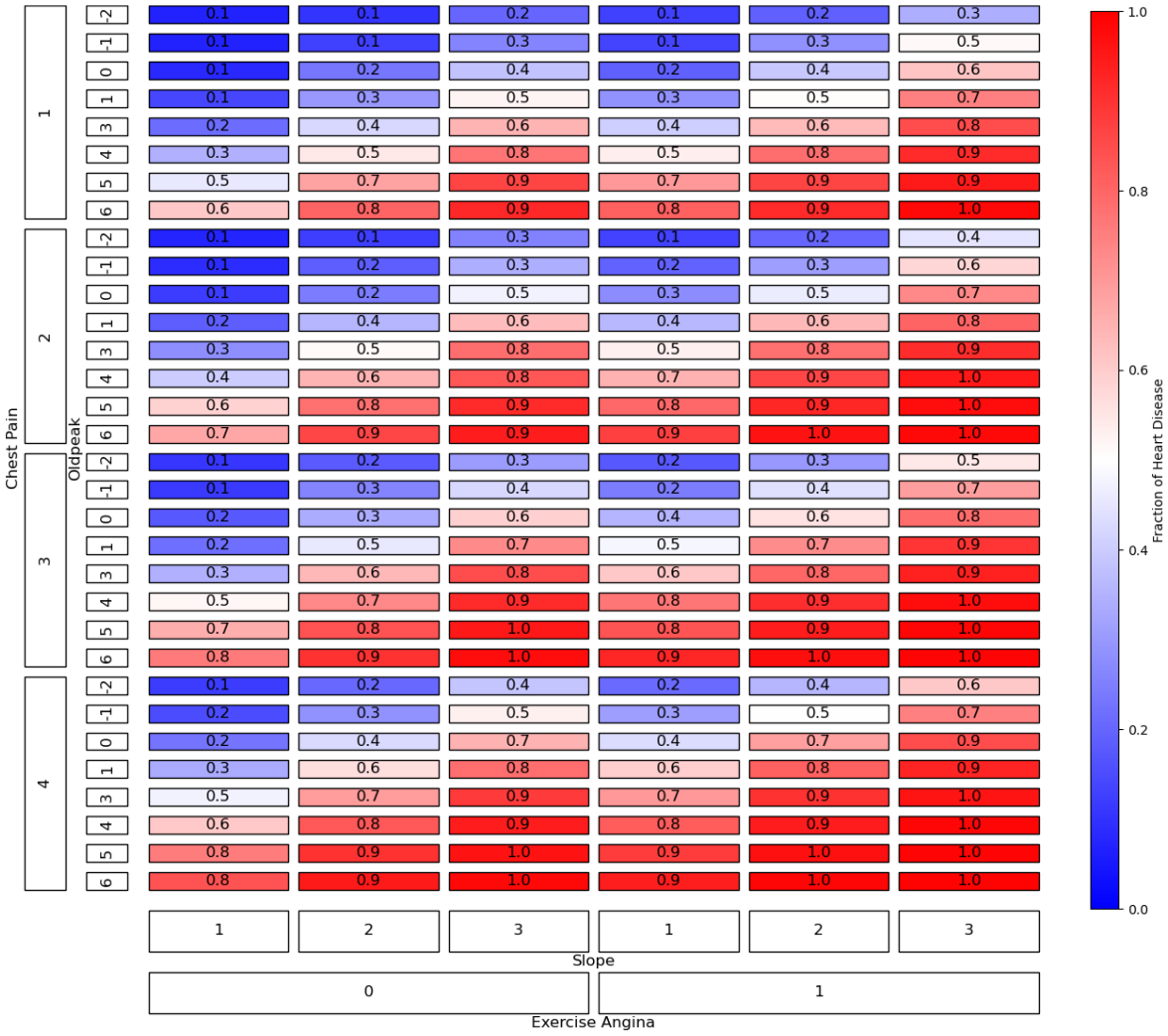


Abbildung A.28.: This figure illustrates the patterns learned by the all features neural net from the heart disease dataset based on the calibrations. The depth of the colors represent the fraction of heart disease.

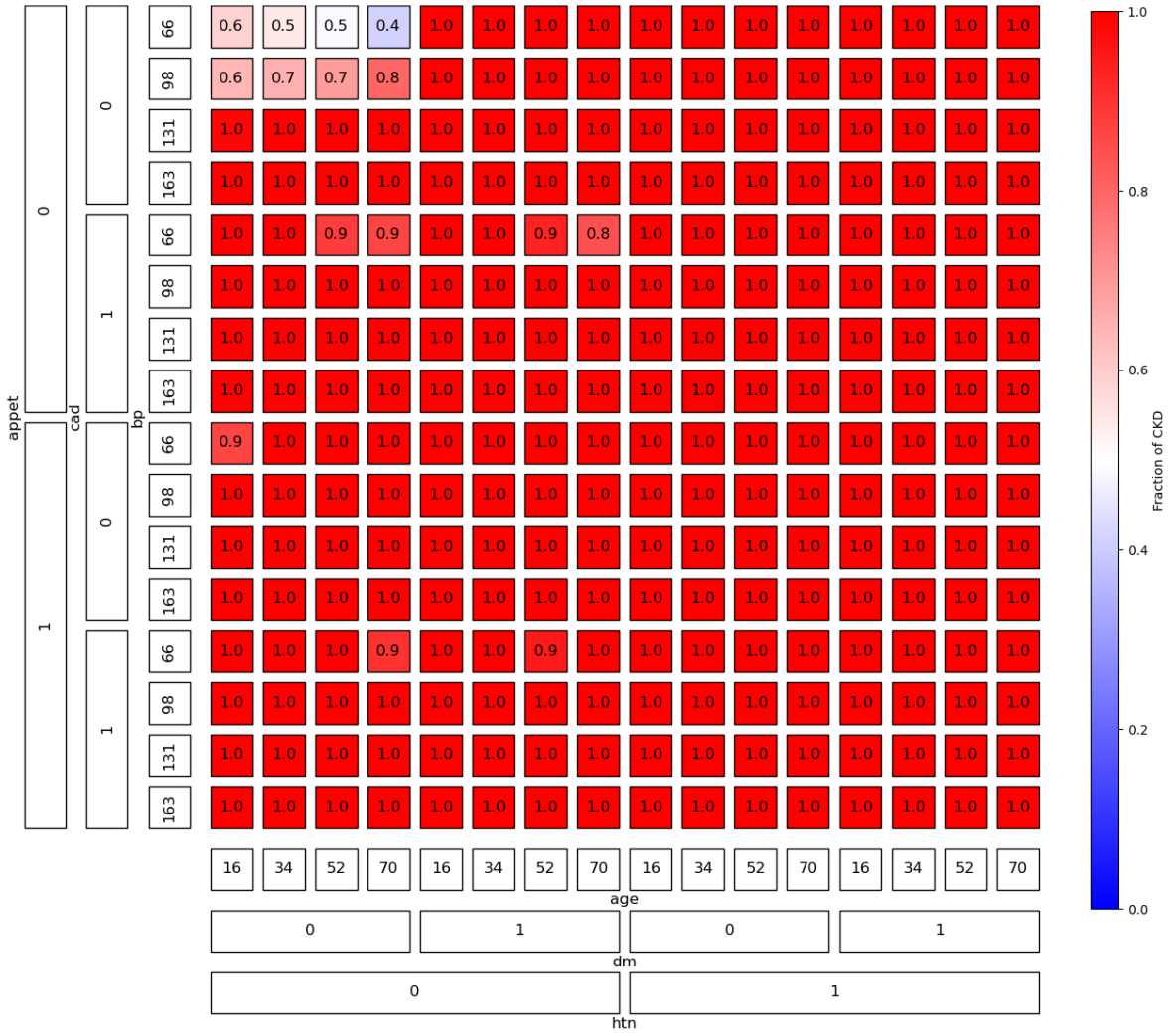


Abbildung A.29.: This figure illustrates the patterns learned by the non-invasive neural net from the chronic kidney disease dataset based on the predictions. The depth of the colors represent the fraction of chronic kidney disease.

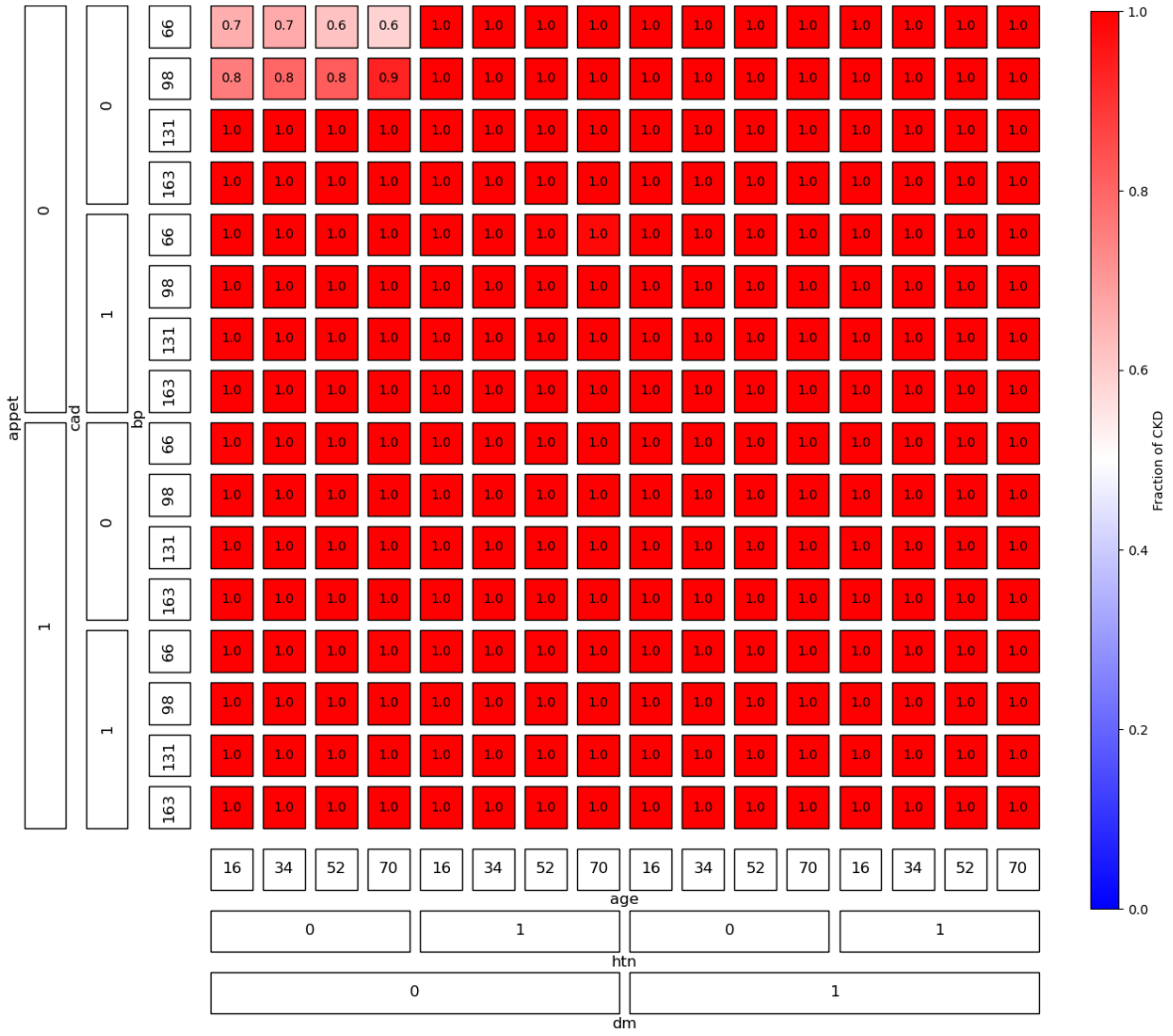


Abbildung A.30.: This figure illustrates the patterns learned by the non-invasive neural net from the chronic kidney disease dataset based on the calibrations. The depth of the colors represent the fraction of chronic kidney disease.

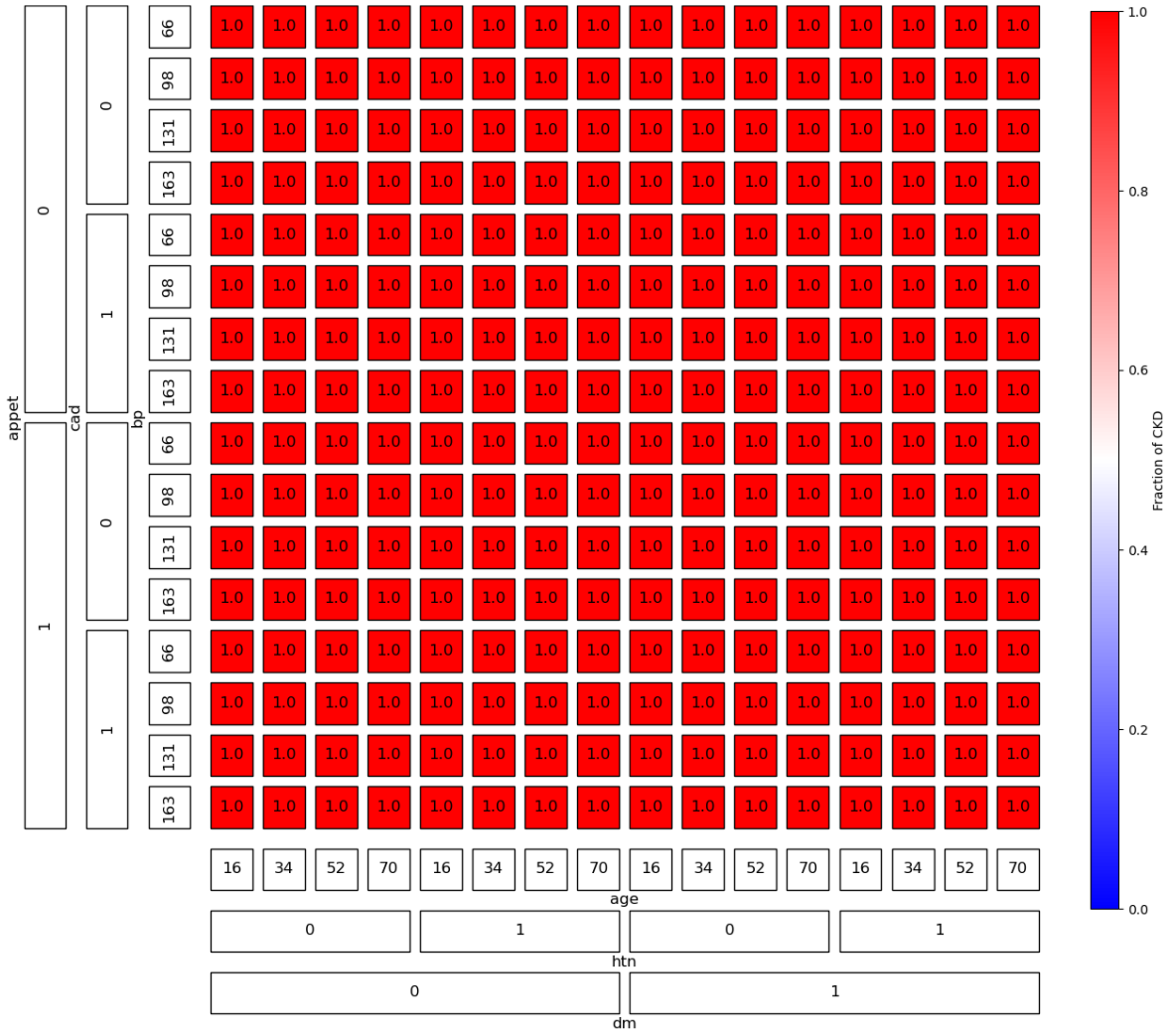


Abbildung A.31.: This figure illustrates the patterns learned by the all features neural net from the chronic kidney disease dataset based on the predictions. The depth of the colors represent the fraction of chronic kidney disease.

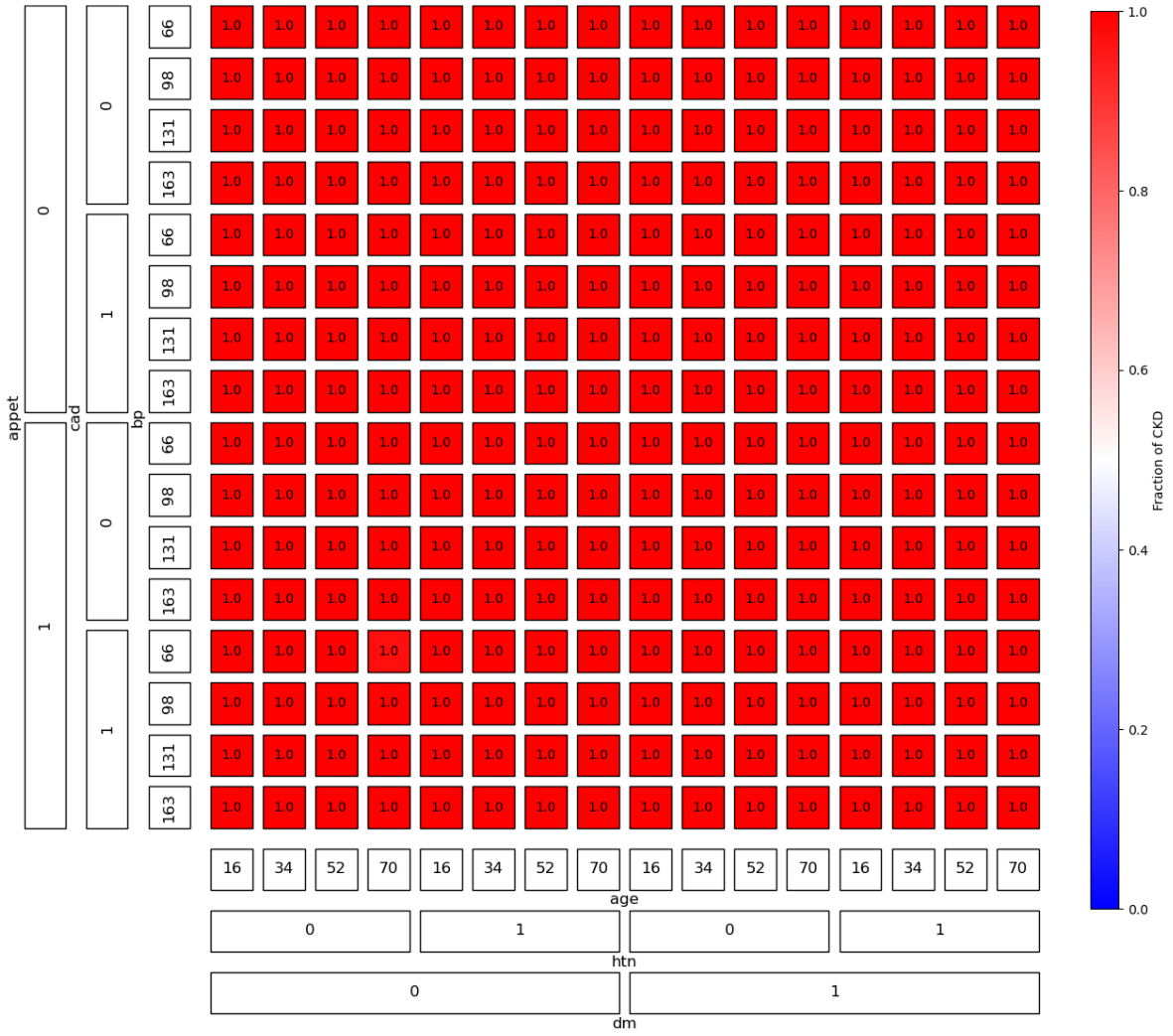


Abbildung A.32.: This figure illustrates the patterns learned by the all features neural net from the chronic kidney disease dataset based on the calibrations. The depth of the colors represent the fraction of chronic kidney disease.

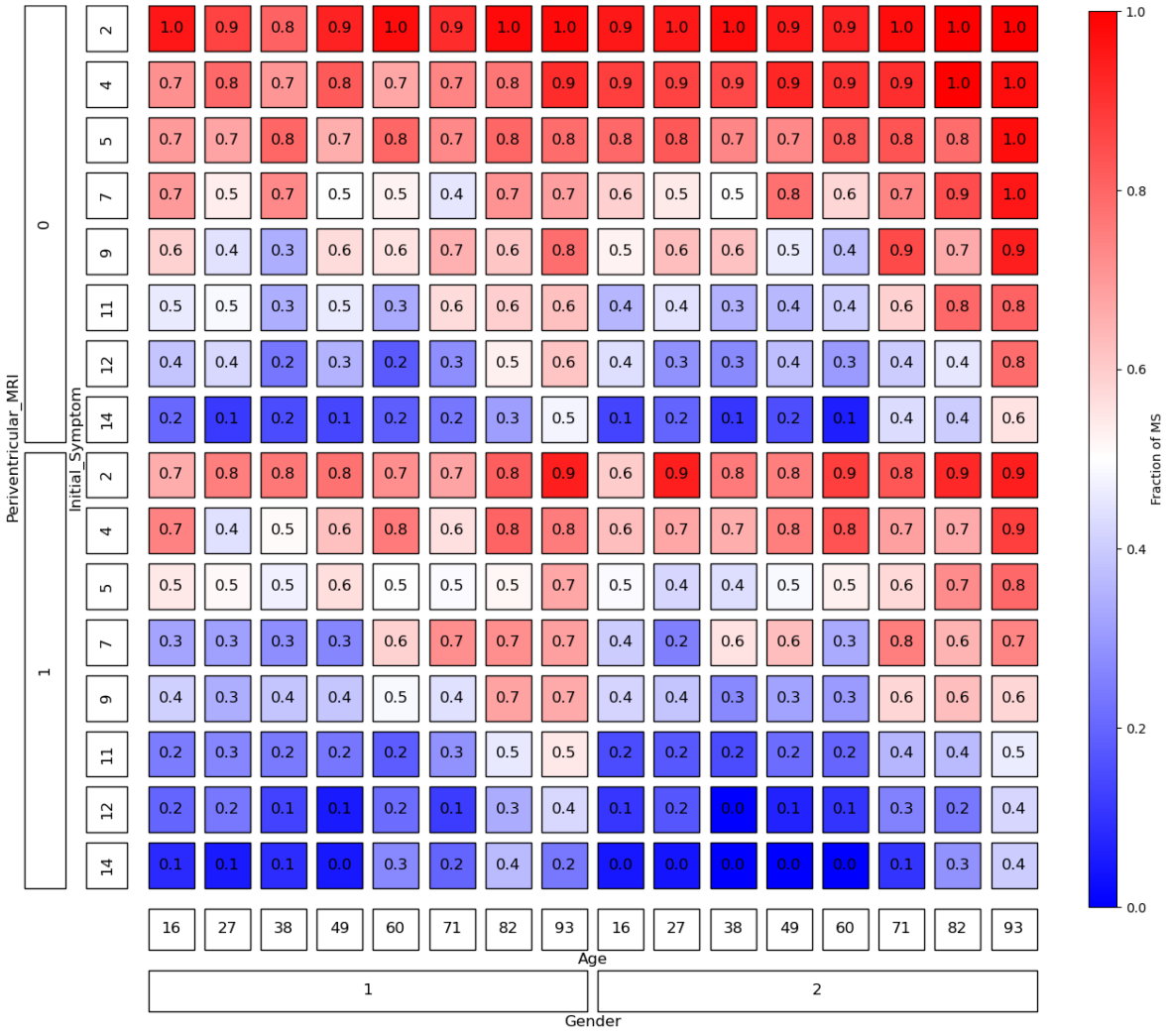


Abbildung A.33.: This figure illustrates the patterns learned by the non-invasive neural net from the multiple sclerosis dataset based on the predictions. The depth of the colors represent the fraction of multiple sclerosis.

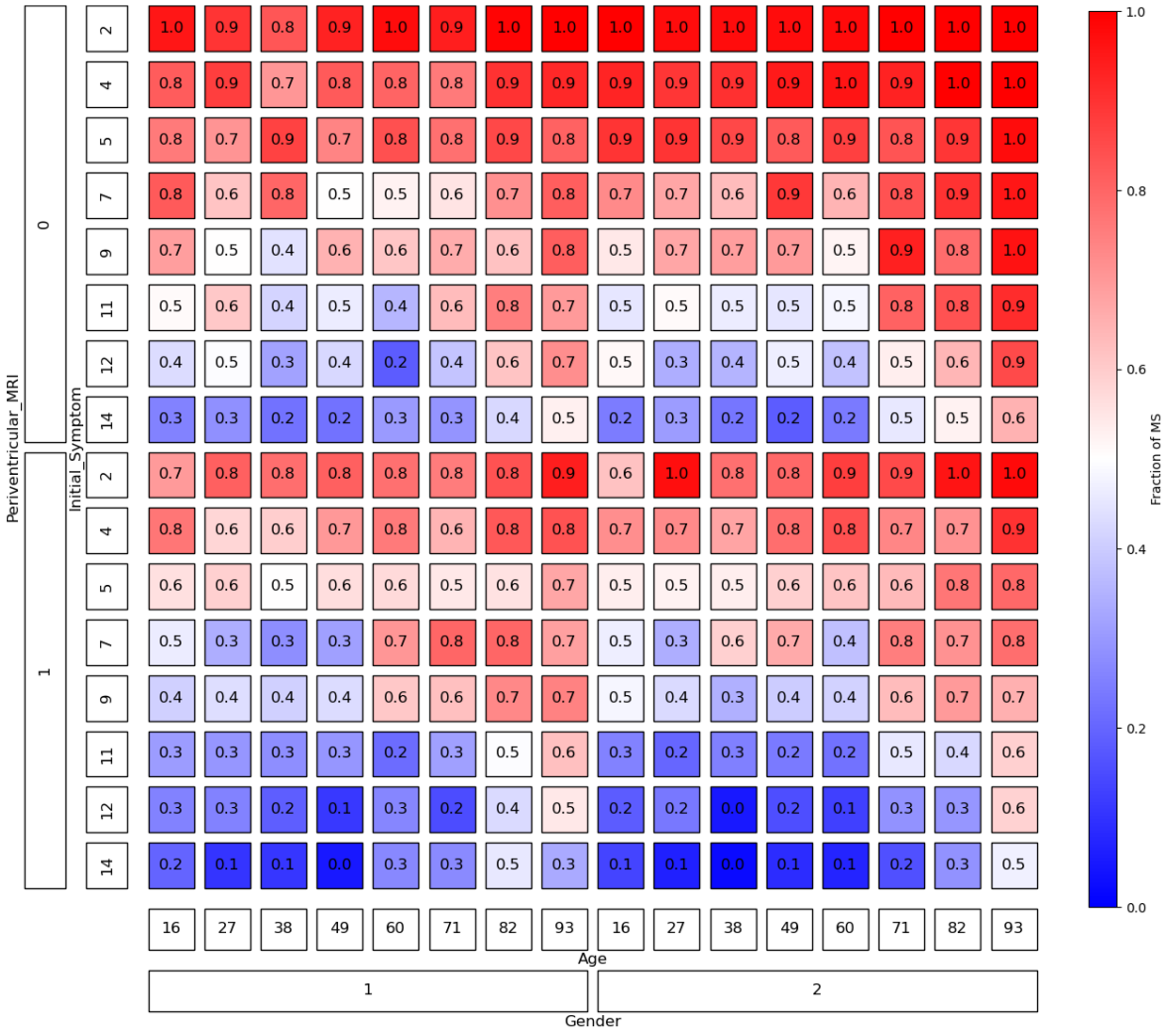


Abbildung A.34.: This figure illustrates the patterns learned by the non-invasive neural net from the multiple sclerosis dataset based on the calibrations. The depth of the colors represent the fraction of multiple sclerosis.

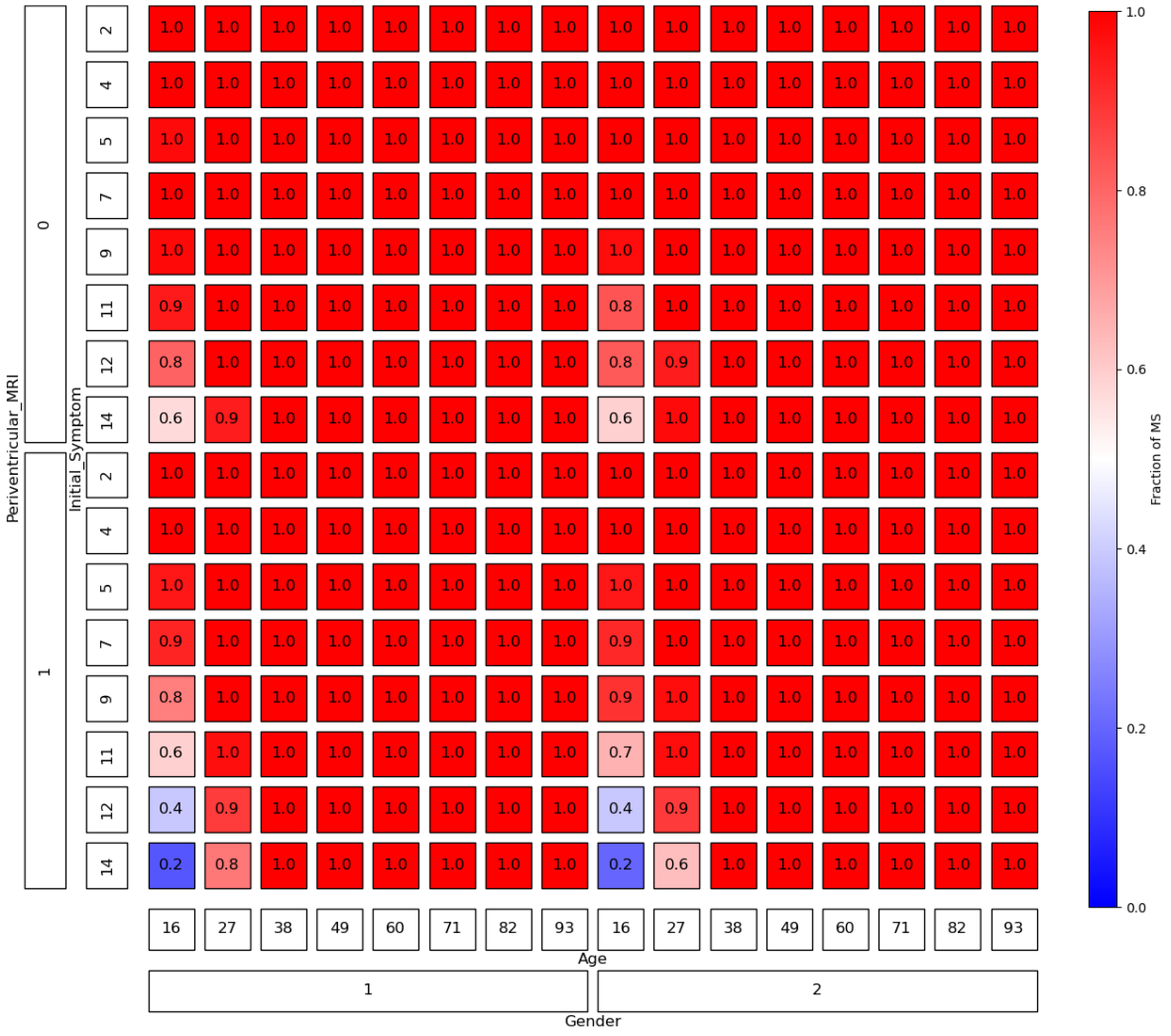


Abbildung A.35.: This figure illustrates the patterns learned by the all features neural net from the multiple sclerosis dataset based on the predictions. The depth of the colors represent the fraction of multiple sclerosis.

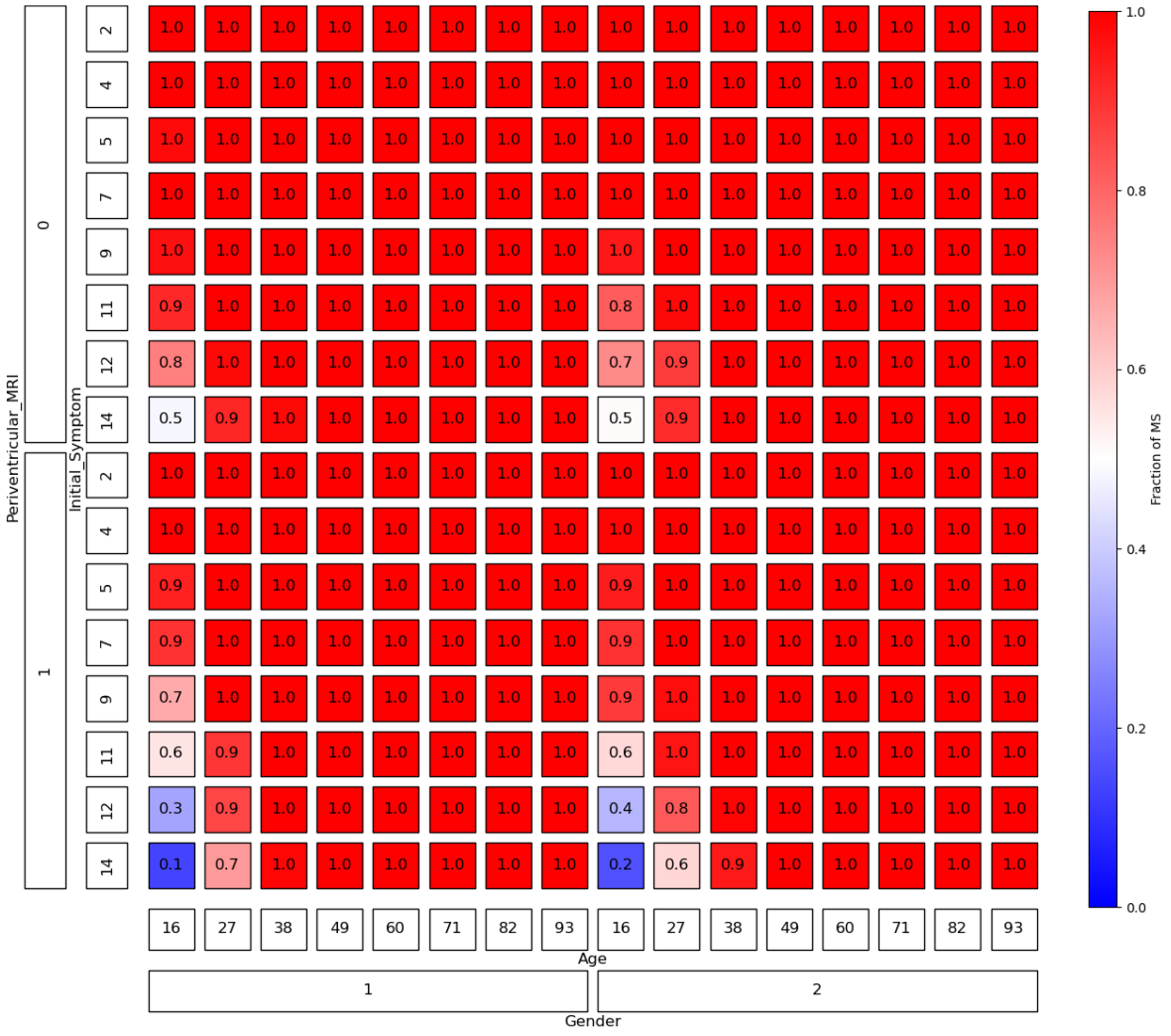


Abbildung A.36.: This figure illustrates the patterns learned by the all features neural net from the multiple sclerosis dataset based on the calibrations. The depth of the colors represent the fraction of multiple sclerosis.

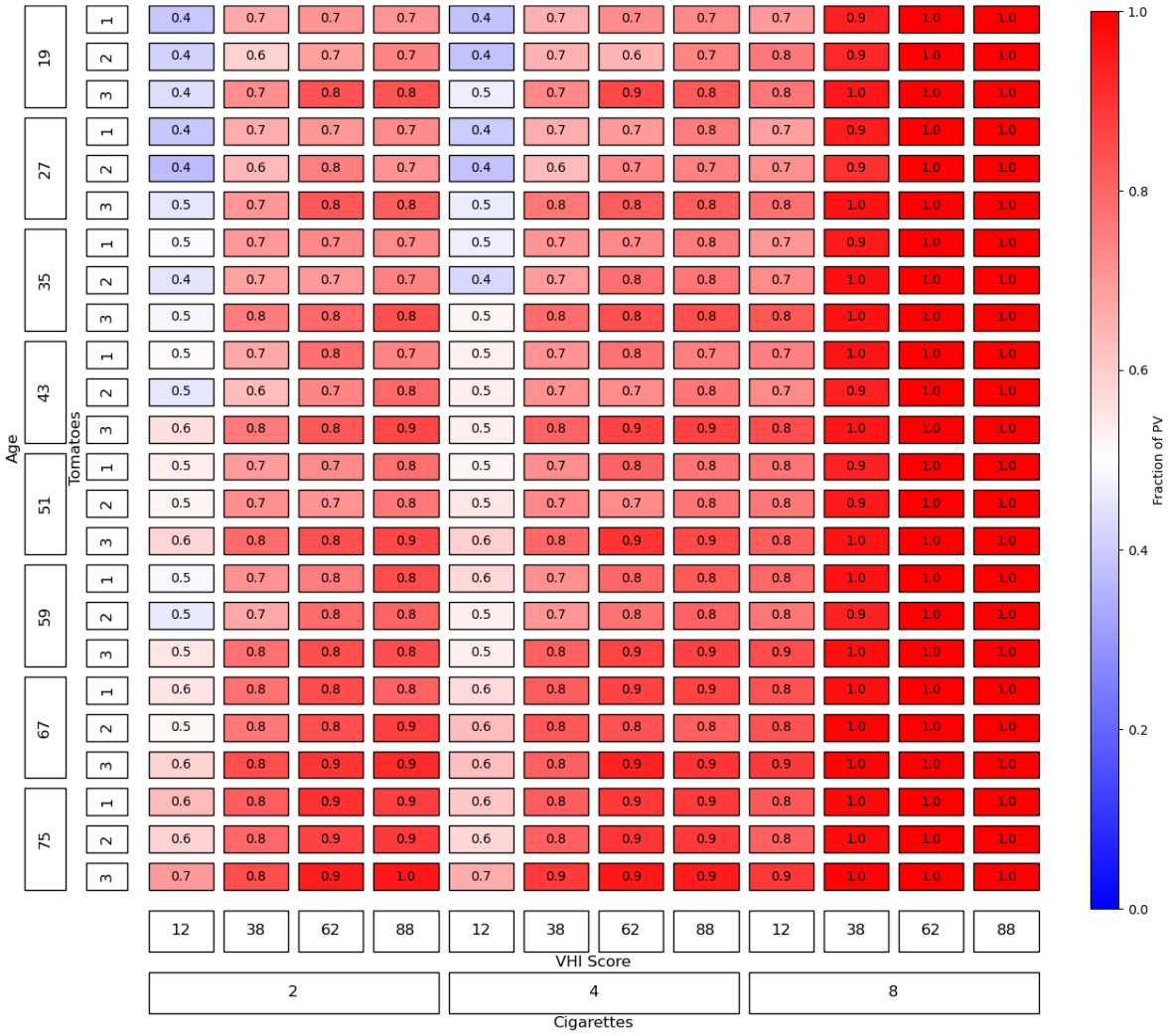


Abbildung A.37.: This figure illustrates the pivot table from the random forest trained on the pathological voice dataset based on its predictions. The depth of the colors represent the fraction of pathological voice.

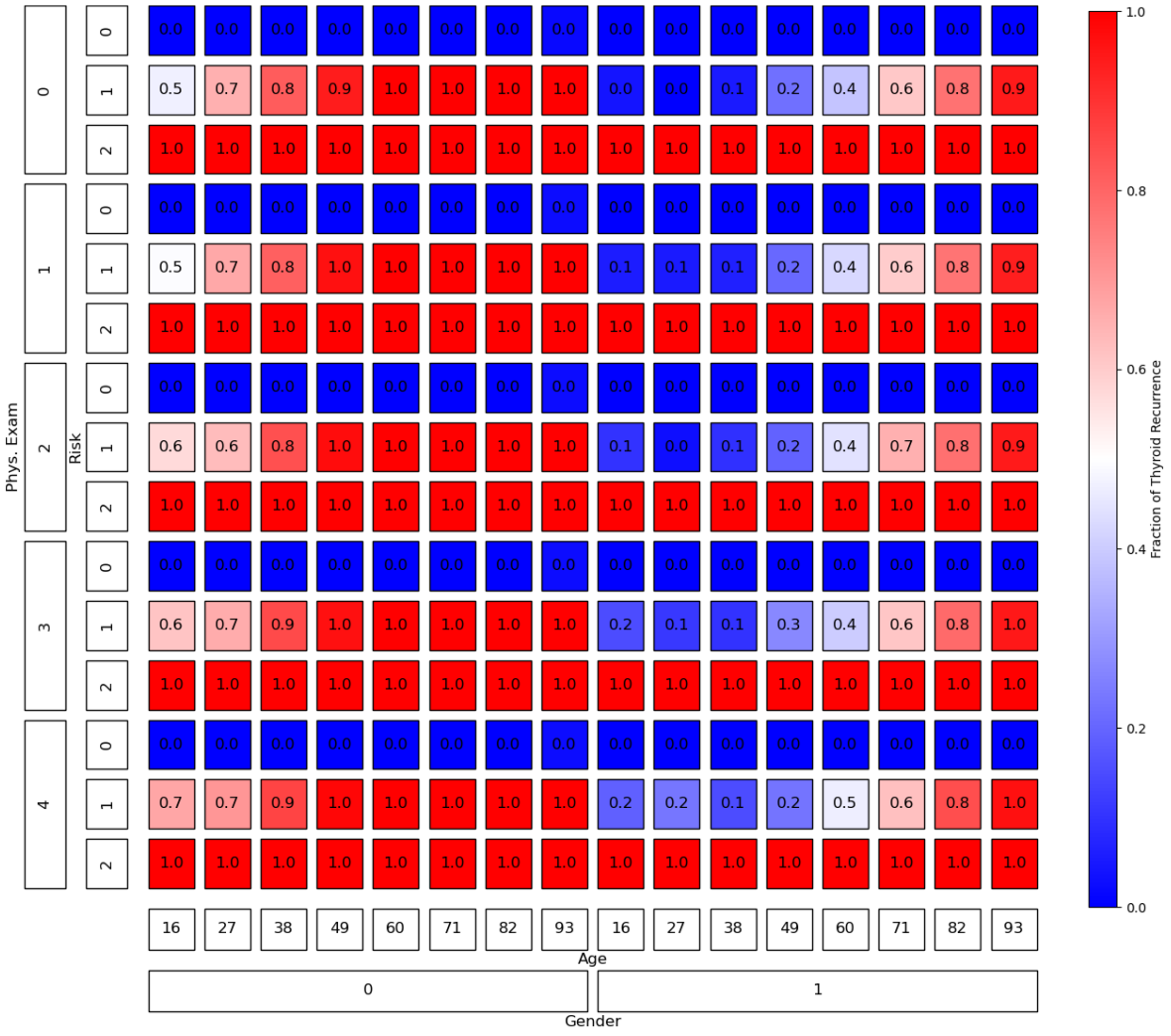


Abbildung A.38.: This figure illustrates the patterns learned by the non-invasive neural net from the differentiated thyroid cancer recurrence dataset based on the predictions. The depth of the colors represent the fraction of thyroid cancer recurrence.

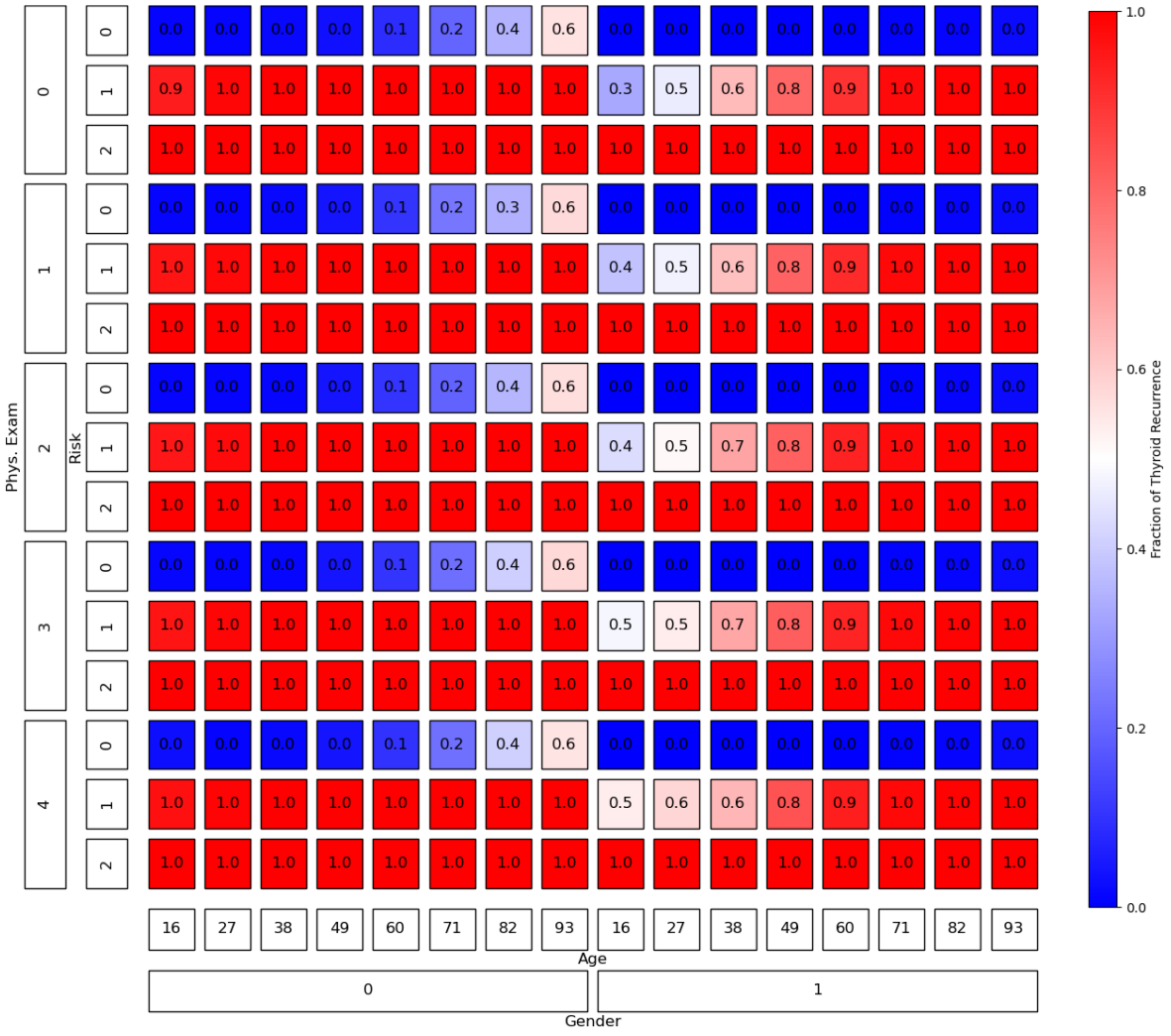


Abbildung A.39.: This figure illustrates the patterns learned by the non-invasive neural net from the differentiated thyroid cancer recurrence dataset based on the calibrations. The depth of the colors represent the fraction of thyroid cancer recurrence.

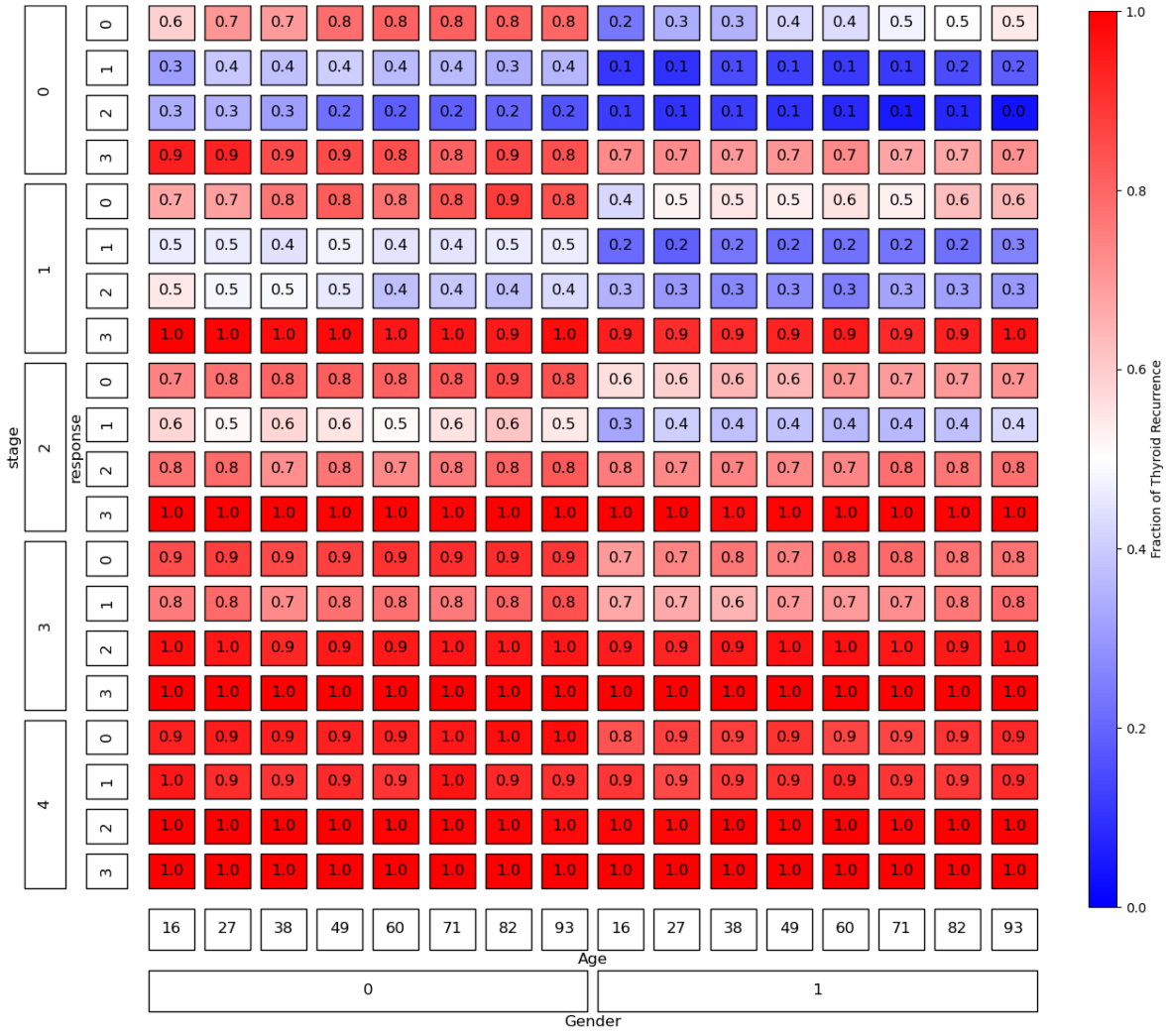


Abbildung A.40.: This figure illustrates the patterns learned by the all features neural net from the differentiated thyroid cancer recurrence dataset based on the predictions. The depth of the colors represent the fraction of thyroid cancer recurrence.

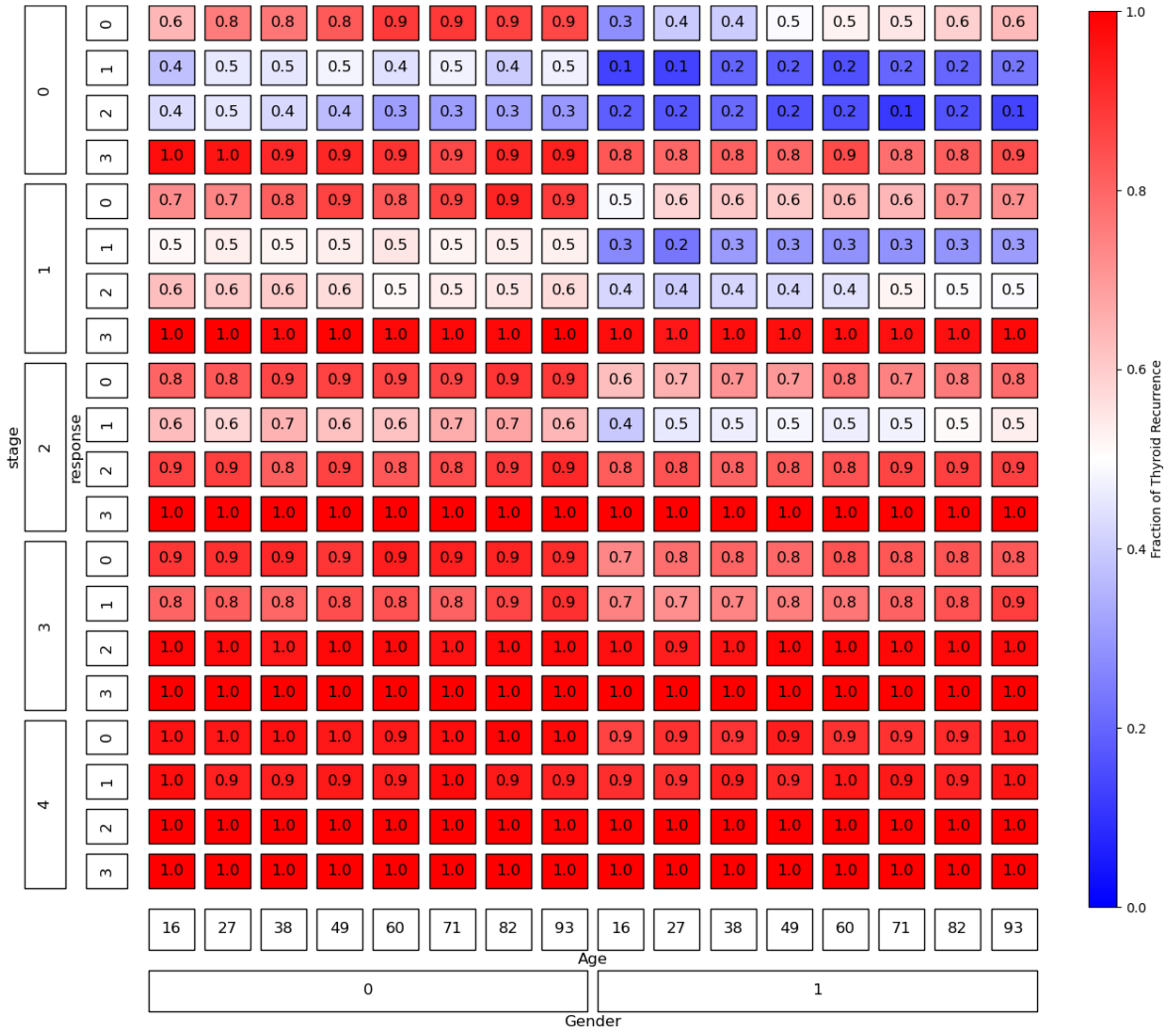


Abbildung A.41.: This figure illustrates the patterns learned by theall features neural net from the differentiated thyroid cancer recurrence dataset based on the calibrations. The depth of the colors represent the fraction of thyroid cancer recurrence.

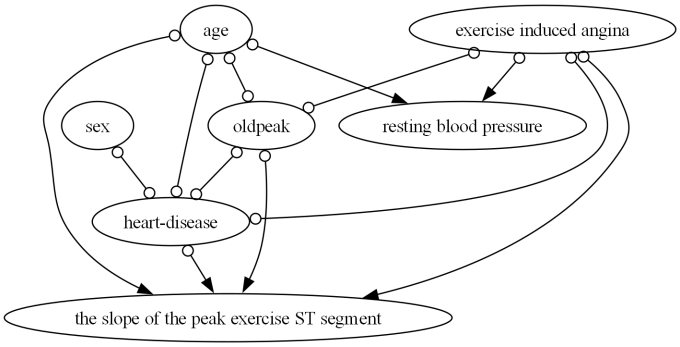


Abbildung A.42.: This figure shows the PAG from the causal discovery on the heart disease dataset. The FCI algorithm made use of the Fisher's Z conditional independence test.

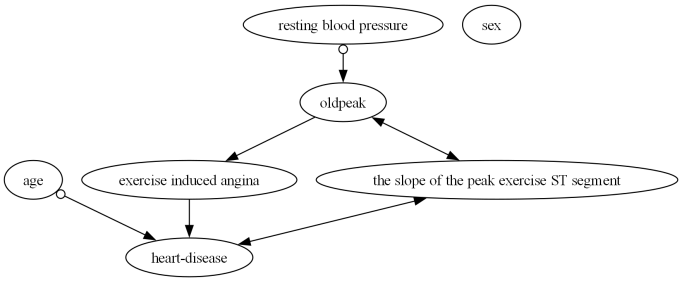


Abbildung A.43.: This figure shows the PAG from the causal discovery on the heart disease dataset. The FCI algorithm made use of the Chi-squared conditional independence test.

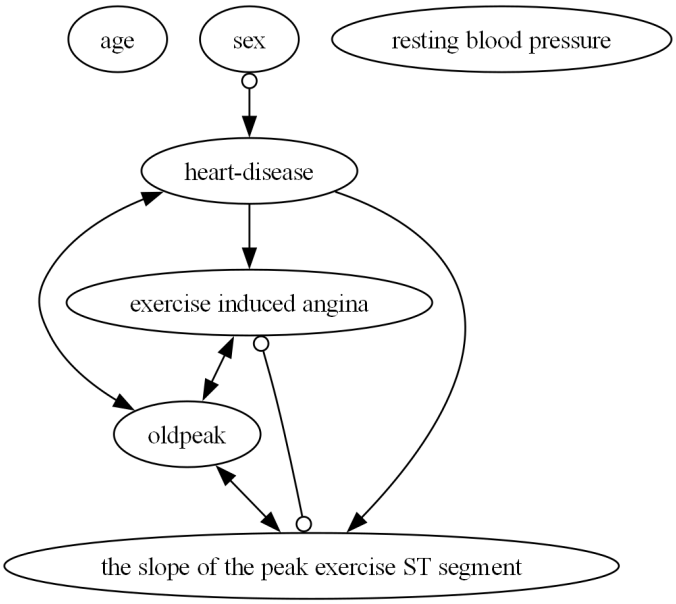


Abbildung A.44.: This figure shows the PAG from the causal discovery on the heart disease dataset. The FCI algorithm made use of the G-squared conditional independence test.

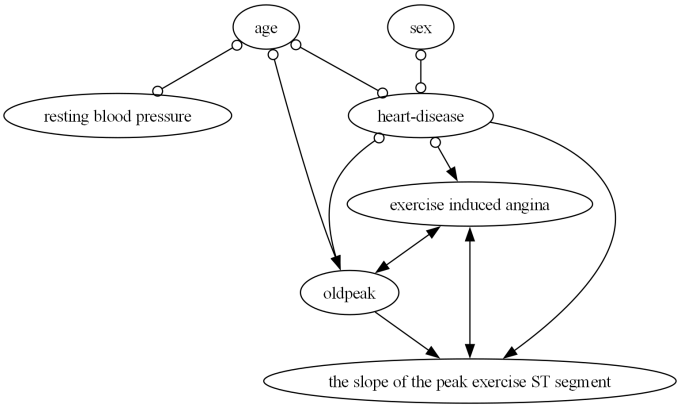


Abbildung A.45.: This figure shows the PAG from the causal discovery on the heart disease dataset. The FCI algorithm made use of the Kernel-based conditional independence test.

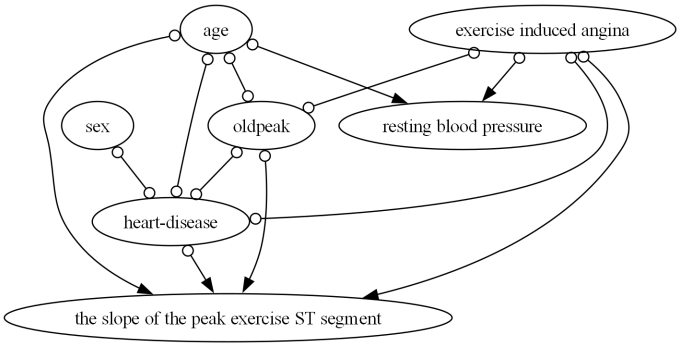


Abbildung A.46.: This figure shows the PAG from the causal discovery on the heart disease dataset. The FCI algorithm made use of the Missing-value Fisher's Z conditional independence test.

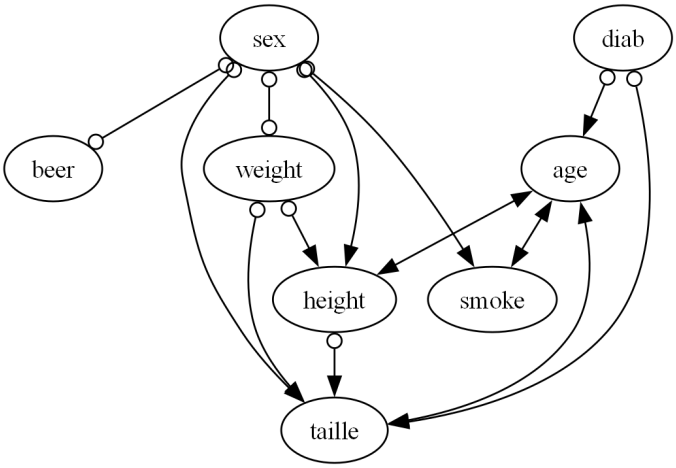


Abbildung A.47.: This figure shows the PAG from the causal discovery on the diabetes disease dataset. The FCI algorithm made use of the Fisher's Z conditional independence test.