UNIVERSITY OF TURKU

Privacy Risks of Explainable Artificial Intelligence (XAI) in Healthcare

An Empirical Study Using Machine Learning Models and Post-hoc Explainability Methods

Cyber Security

Master's Degree Programme in Information and Communication Technology

Department of Computing, Faculty of Technology

Master of Science in Technology Thesis

Author:

Olli Saaristo

Supervisors:

Jouni Isoaho

Kaitai Liang

28.5.2026

Turku

Master of Science in Technology Thesis
Department of Computing, Faculty of Technology
University of Turku

Subject: Cyber Security
Programme: Master's Degree Programme in Information and Communication Technology
Author(s): Olli Saaristo
Title: Cybersecurity and Privacy Risks of Explainable Artificial Intelligence (XAI) in Healthcare
Supervisor(s): Prof. Jouni Isoaho, Dr. Kaitai Liang
Number of pages: 66 pages, 5 appendix pages
Date: 28.5.2026

This thesis examines the cybersecurity and privacy implications of Explainable Artificial Intelligence (XAI) methods in selected Machine Learning (ML) models. The study focuses on whether post-hoc explainability methods can expose sensitive information from healthcare datasets when applied to machine learning models.

The study was conducted using a modified synthetic healthcare dataset. The dataset consisted of 1,000 synthetic patient records with intentionally modified values to simulate privacy-related anomalies. Three machine learning models were implemented using the Scikit-learn library in JupyterLab: Decision Trees (DTs), Support Vector Machines (SVMs), and K-Nearest Neighbors (KNN). The explainability methods applied in the study were Local Interpretable Model-agnostic Explanations (LIME), Shapley Additive exPlanations (SHAP), and Explain Like I'm Five (ELI5).

The results indicate that XAI methods improve the transparency and interpretability of ML models but may also increase the exposure of sensitive information by highlighting influential patient features and unusual data patterns. Decision Trees worked as white-box models and were the most directly interpretable models, whereas SVMs and KNN worked as black-box models and required additional XAI methods for interpretability.

However, the applied privacy-preserving techniques may influence model behavior and should be carefully considered when interpreting the results. Increased privacy protection may be associated with reduced model accuracy and a potential increase in bias.

The study concludes that explainability and privacy must be carefully balanced when using AI systems in healthcare environments. European Union regulations and global legislation emphasize that patient information must be handled with appropriate privacy and security measures.

Key words: Artificial Intelligence (AI), Explainable AI (XAI), Cybersecurity, Privacy, Healthcare, Trustworthy AI, Regulations

Acknowledgements

Vapaan kansan lahja vapaalle tieteelle

I would like to express my sincere gratitude to the lecturers in cybersecurity and business at the University of Turku for their inspiring teaching and support throughout my studies. I would also like to thank my supervisors for their guidance and valuable feedback during the thesis process. In addition, I am grateful to the university staff and catering services, who contribute to making academic studies possible.

I would also like to acknowledge Finnish society for providing access to high-quality higher education, as without its support, none of this would have been possible.

I am incredibly fortunate to have the backing of my family throughout my academic pursuit. I would especially like to thank my mother for her continuous support and encouragement. She has taught me that everything is achievable and that dedicating oneself to education is a genuinely rewarding path.

Table of contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Definition
4IR	the Fourth Industrial Revolution
AI	artificial intelligence
AIA	Artificial Intelligence Act (EU)
CIA	confidentiality, integrity, and availability triad
DL	deep learning
DTs	decision trees
ELI5	explain like I'm five (XAI Python library)
EU	European Union
GDPR	General Data Protection Regulation
GenAI	generative artificial intelligence
HIPAA	Health Insurance Portability and Accountability Act (US)
ICT	information and communication technologies
IML	interpretable machine learning
IoT	Internet of Things
IT	information technology
KNN	k-nearest neighbors
LIME	local interpretable model-agnostic explanations
LLMs	large language models
ML	machine learning
NIS2	Network and Information Security Directive (EU)

PII	personally identifiable information (US)
SHAP	Shapley additive explanations
SVMs	support vector machines
US	United States
XAI	explainable artificial intelligence
X-PRISM	Explainable Privacy-Preserving Intelligent System for Monitoring

The use of AI

Artificial intelligence (AI) tools, large language models ChatGPT (OpenAI, GPT-5 series) and Gemini (Google), were used during the writing process of this thesis as supportive tools. These tools were primarily used for grammar checking and improving the clarity and readability of selected text sections. In addition, they were used as assistants in creating Python and Scikit-learn scripts. The AI tools helped with debugging errors and programming-related questions, particularly in cases where changes in the Scikit-learn machine learning library affected the program's functionality. For example, some function calls related to decision trees were outdated and required updates to work correctly.

These tools were also used for structuring the thesis, formatting mathematical formulas in Microsoft Word using LaTeX syntax, brainstorming, and discussing ideas.

No AI-generated content was used as such, and all ideas and output generated by the AI tools were always reviewed and validated by the author. All critical analysis and research were done by the author.

1 Introduction

The industrial timeline can be divided into four revolutions. Each Industrial Revolution has fundamentally transformed how we work and how the world is shaped. The First Industrial Revolution of 1765 started mechanization with the help of steam engines. By 1870, the transition into the Second Industrial Revolution led to the discovery of electricity, gas, and oil, resulting in inventions such as: automobiles, airplanes, telegraph, and telephone. The digital landscape shifted significantly with the Third Industrial Revolution of 1969, which integrated electronics, advanced computing, and global telecommunications into mainstream industry. Most recently, the mid-2010s marked the onset of the Fourth Industrial Revolution (4IR), which redefines global industry through ubiquitous connectivity, intelligent automation, and next-generation computing [1] [2].

We are currently living in the 4IR era and witnessing the rise of artificial intelligence (AI) systems that are shaping and speeding up the development of the world of information technology (IT). There are multiple benefits that AI can offer across various industries and applications, such as automation of repetitive tasks, faster processing of data, assist with decision making, and fewer human errors [3].

AI, robotics, and Internet of Things (IoT), among other things, have made a huge impact on 4IR. The healthcare sector is rapidly adopting these technologies to improve diagnostics, treatment, and overall patient care. AI can assist with tasks that traditionally require human intelligence, such as to solve problems in advanced healthcare diagnostics [4].

While AI has remarkable potential to shape our lives, it is important to understand what it is – and what it is not – and what problems arise when it is implemented in everyday life. In particular, machine learning and generative AI have been a hot topic as trustworthiness, cybersecurity, and privacy issues has gained a lot of attention [5] [6].

The focus of this thesis is on the and privacy risks of explainable AI (XAI) in the healthcare sector and on related legislative perspectives within the European Union (EU).

The most frequently used abbreviations are listed in the List of Abbreviations at the beginning of this thesis. Less frequent terms are defined upon their first occurrence in the text to enhance readability but are not used later on (e.g., the European Cybersecurity Certification Framework, ECCF). Additionally, a distinction is made between proper nouns (capitalized) and common nouns (lowercase).

1.1 Research Background

Major challenges with AI include the lack of explainability, how the conclusion was reached. In aviation, flight recorders, commonly referred to as “black boxes”, enhance transparency by recording all relevant flight data for later analysis, but a “black box” in AI systems refers to a model whose deductions are opaque to human observers. In opaque black-box systems errors are difficult to detect and debugging the root cause may be impossible. In the healthcare sector it is important to understand how the AI model works and generates predictions, to avoid catastrophic failures. Black-box models also raise concerns about potential AI model biases, for example, if the training data is insufficient [7].

Explainable artificial intelligence (XAI; not to be confused with the social media platform X’s, formerly known as Twitter, AI project “xAI”) aims to make AI decision-making transparent and interpretable. Different XAI methods add trustworthiness to AI solutions, making them more viable and ubiquitous in future healthcare environments. In 2019, the EU defined key guidelines for trustworthy AI, emphasizing that AI systems should be lawful, ethical, and robust. These three principles are reflected in seven requirements that emphasizes, among other things, meaningful human control, technical robustness and safety, strong privacy and data governance, openness and transparency, fairness and inclusion, positive societal and environmental impact, and clear mechanisms for accountability [8]. In 2024, the EU adopted the Artificial Intelligence Act (AIA), which introduces regulatory requirements for the safe and responsible use of AI [9]. AIA is discussed in more detail in Chapter 3.

1.2 Research Questions

The following research questions (RQ) will guide this thesis:

- RQ1: What are the key cybersecurity and privacy risks associated with the use of AI in healthcare systems?
- RQ2: How do XAI methods (such as LIME, SHAP, ELI5) influence the exposure of sensitive information in healthcare datasets?
- RQ3: How do dataset characteristics (e.g., unique values, feature distribution) influence model explainability and potential privacy risks?

1.3 Research Objectives

The main objective of this thesis is to investigate the cybersecurity and privacy risks associated with AI systems in the healthcare sector and to examine how XAI methods can help address these challenges. To answer the research questions presented in Section 1.2, the study aims to:

- Identify key cybersecurity and privacy risks introduced by AI-based healthcare systems (RQ1).
- Compare selected XAI methods (LIME, SHAP, and ELI5) in terms of their ability to explain model predictions (RQ2).
- Evaluate how dataset characteristics, such as unique values, affect model performance and stability of explanations (RQ3).
- Study whether XAI methods can expose sensitive patient data or introduce additional privacy risks (RQ2, RQ3).

1.4 Scope and Limitations

This thesis seeks to identify what new cyberthreats and privacy issues are created by AI and how XAI can be part of the future solutions in the healthcare sector within the European Union (EU). The EU's General Data Protection Regulation (GDPR) creates a strict framework for how sensitive data should be handled, and similar regulatory approaches are emerging in other regions. In addition, more recent regulations and directives, such as the Network and Information Security Directive (NIS2) and the AI Act (AIA), have come into effect.

The scope of this thesis is limited to machine learning-based healthcare applications. Deep learning methods are also briefly discussed. The analysis focuses on selected explainability methods (LIME, SHAP, and ELI5) and does not cover all existing XAI techniques.

This thesis focuses on cybersecurity and privacy perspectives within the EU, while other regulatory frameworks are briefly mentioned but not analyzed in depth.

Philosophical issues such as "What is intelligence?" and "Can machines think?" are outside the scope of this thesis. Additionally, ethical issues, such as job losses and environmental

impacts, are excluded. Lastly, AI governance is not discussed other than legislative perspectives.

1.5 Thesis Structure

This thesis is structured into seven main chapters.

- Chapter 1: Introduction – The first chapter presents the research problem, research questions, research objectives, and scope and limitations of the study.
- Chapter 2: Background – The second chapter introduces the theoretical background and key concepts related to AI and XAI.
- Chapter 3: Cybersecurity and Privacy – The third chapter discusses cybersecurity, privacy, and legislative considerations related to AI and XAI, with an emphasis on the European Union perspective.
- Chapter 4: Methodology – The fourth chapter describes research design, dataset preparation, and methods used in conducting the research.
- Chapter 5: Results – The fifth chapter presents the results of used machine learning models and XAI methods.
- Chapter 6: Discussion – The sixth chapter interprets the findings and discusses their implications.
- Chapter 7: Conclusion – This final chapter provides a synthesis of the primary insights and contributions established in this work, while also outlining potential avenues for subsequent studies.

2 Background

This chapter gives an overview of AI fundamentals and common use cases, emphasizing machine learning, deep learning, and XAI methods. Some AI approaches are outside of the scope of this thesis and are therefore left out, while others are only briefly mentioned, such as diffusion models. AI philosophy, AI governance, and ethics are also outside of the scope.

The background literature for this study was collected from academic databases, primarily Google Scholar and IEEE Xplore. Keywords such as “cybersecurity”, “healthcare”, “explainable artificial intelligence”, and “XAI SHAP”, were used to find relevant documentation and scientific articles. Additionally, government reports, industry publications, and other credible online sources were used.

2.1 Artificial Intelligence

Alan Turing (1912–1954), a British scientist, became famous for the concept of the Turing Machine in 1936 [10] and later the Turing Test in 1950. He is commonly referred to as one of the early pioneers of artificial intelligence. During the Second World War, he played a key role in deciphering the Enigma cipher machine used by Germany, demonstrating early practical applications of computational thinking [3].

John McCarthy (1927–2011), an American scientist, introduced the term “artificial intelligence” in 1956. He described AI as a scientific and engineering field which essentially focuses on building intelligent machines. In particular, he referred to computer programs and emphasized that such systems do not need to replicate the exact mechanisms of human cognition [11].

There is no universally agreed definition of AI, and instead it functions as an umbrella term covering various techniques that allow computers and machines to mimic human-like reasoning and decision-making. The definition of AI has evolved over time, and many early computer programs would no longer be considered AI by modern standards. At present, AI is often seen as a set of machine learning algorithms based on neural networks, but in the future this definition may not be satisfactory anymore [12]. In this thesis, “AI” primarily refers to machine learning and related technologies.

Rule-based systems represent one of the simplest forms of AI: they apply predefined if-then-else rules to make decisions without human intervention. A simple example is a thermostat

that adjusts heating or cooling automatically when the room temperature crosses specified thresholds [13].

AI has evolved over decades from early rule-based systems to machine learning approaches, and more recently to advanced deep learning systems based on neural networks. Figure 1 illustrates how AI has developed over more than 70 years, from the 1950s to the present day. A simple way to categorize AI is to divide it into subfields: early rule-based AI in the 1950s was followed by machine learning (ML) in the 1980s, then deep learning (DL) in the 2010s, and more recently generative AI (GenAI) in the 2020s, which is based on deep learning techniques [3]. It is important to distinguish the subfields in order to better understand how different AI tools operate and what they can do.

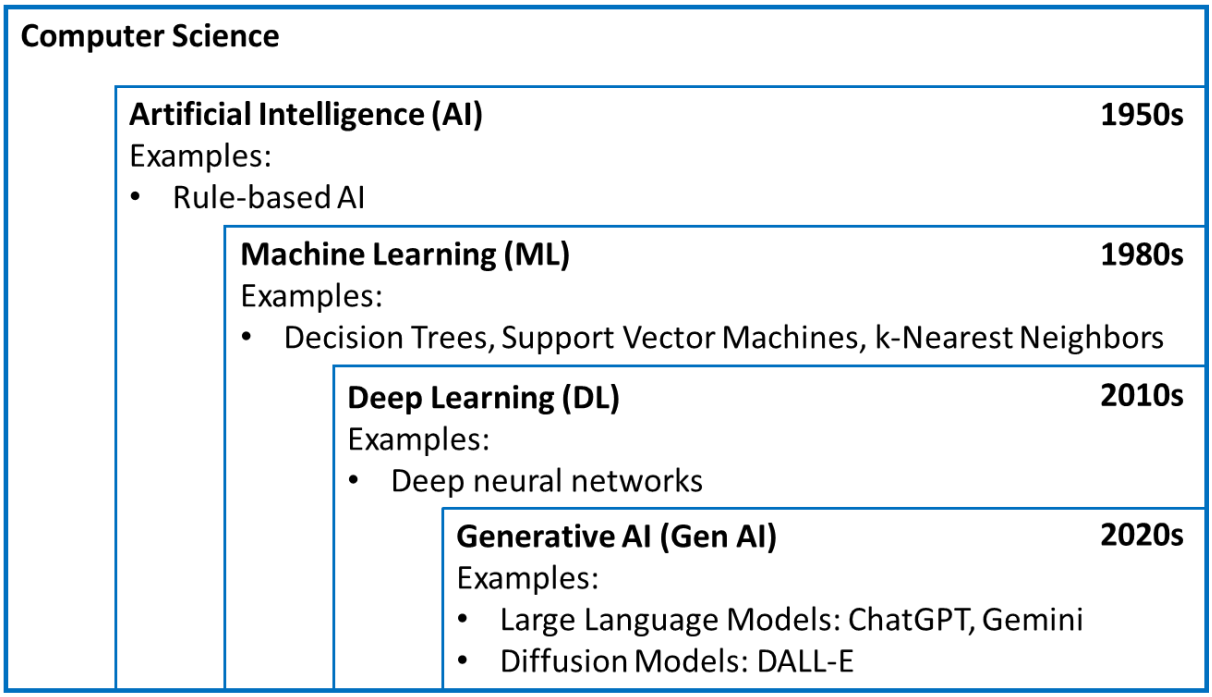


Figure 1. Nested structure of AI concepts, illustrating the classification of ML, DL, and GenAI within computer science.

It is worth noting that all of today's artificial intelligences are weak AI (sometimes referred to as artificial narrow intelligence (ANI) or narrow AI), meaning they are trained to perform a specific task [14]. Strong AI (sometimes referred to as artificial general intelligence (AGI) or general AI) would be able to perform any intellectual task that a human can; however, this kind of technology does not yet exist [15]. Another conceptual AI system is super AI

(sometimes referred to as artificial superintelligence (ASI)), which would exceed human intelligence in all areas [14].

GenAI can produce novel content, such as text-based, sound-based, or image-based content. Large Language Models (LLMs) generate human-like text and can assist with various reasoning and problem-solving tasks. Notable examples of LLMs include the United Stated (US) based OpenAI's ChatGPT, Google's Gemini, Microsoft's Copilot, and France-based Mistral AI's Mistral. In healthcare, the use of LLMs has increased due to their ease of use and fast and understandable analysis of patient data, decision support, and information retrieval.

2.2 Machine Learning

Machine learning (ML) is a branch of AI in which models are designed to learn patterns from data instead of relying solely on preprogrammed code. The models can adapt through repeated exposure to examples (i.e., experience) and gradually improve at performing the desired tasks. During training, the algorithm is provided with input data and corresponding target outputs and adjusts its internal parameters to enhance performance. This process of adaptation is referred to as training [16]. The field of ML includes many different algorithmic approaches, for example, decision trees, various clustering methods, linear regression, and artificial neural networks [3].

2.2.1 Labeled Data and Unlabeled Data

To use ML tools effectively, it is important to distinguish whether the data is labeled, unlabeled, or both, as it determines what kind of ML method should be used. For example, supervised learning is applied to labeled data, whereas unsupervised learning is used for unlabeled data. Semi-supervised learning can be used if the data is both labeled and unlabeled [17].

Data labeling approaches vary, as there are many ways to label, such as synthetic labeling, programmatic labeling, internal labeling, outsourcing, and crowdsourcing. Internal labeling relies on professional data science experts, but this approach is costly. Synthetic labeling generates new data from preexisting datasets, providing greater accuracy, but this approach requires significant computing resources. Programmatic labeling is an automated data labeling process, but technical problems still require human intervention. Outsourcing can be beneficial for short projects, but costs and project management become more difficult over

time. Crowdsourcing is more cost-effective due to its microtasking capability, but maintaining data quality, quality assurance (QA), and project management can be challenging. One famous example of crowdsourced data labeling is reCAPTCHA, which is used to distinguish bots from humans on the Internet. Users are prompted to identify objects in a picture and these responses are later used for data labeling purposes [18].

2.2.2 Machine Learning Methods

Machine learning methods are commonly divided into four main types: supervised learning, unsupervised learning, reinforcement learning, and semi-supervised learning. This list is not exhaustive and there are many other approaches to machine learning.

- **Supervised learning** is applied when labeled data are available and the objective is to learn a mapping from inputs to outputs. Training examples are paired with target labels, which are usually assigned by humans, and the model uses these examples to learn how to predict outputs for a previously unseen inputs [3]. Common use cases for supervised learning are classification, regression, and object detection tasks. However, creating labeled data can be expensive and time-consuming as human intervention is needed to assign labels to each data point [17].

  Classification tasks can be divided into binary or multi-class. Some common classification algorithms are Naïve Bayes, logistic regression, decision tree classifiers, k-nearest neighbors (KNN), and support vector machines (SVMs). Some common regression algorithms are simple linear regression, lasso regression, and decision trees regressors [19] [20].

  Supervised learning methods are widely used in applications from medical diagnosis and fraud detection to speech, spam detection, and image recognition [20].

- **Unsupervised learning** is used when the data is unlabeled and the output is not known in advance. The algorithms make their own predictions of what the unlabeled and unstructured data represents, and no human intervention is needed. The model is trained to discover intrinsic patterns, correlations and structures [21]. Unsupervised learning can be used for tasks such as finding patterns, groupings and structures without predefined labels. There are three major subsets of unsupervised learning algorithms: clustering, association, and dimensionality reduction [17].

Clustering algorithms partition unlabeled data into groups, based on measures of similarity or distance between data points. They can also be used as predictive tools for anomaly detection, for example by recognizing points that do not clearly belong to any cluster. Common clustering methods include Gaussian mixture models (GMMs) and k-means [19] [20].

Association algorithms are used in large datasets to identify correlations between variables, such as product recommendation engines to recognize which items are frequently purchased in pairs or groups. Common algorithms are Apriori algorithm, Eclat, and dynamic itemset counting (DIC) [19] [20].

Dimensionality reduction techniques compress data into fewer variables while preserving its core information. Common approaches are for example autoencoders [19].

Unsupervised methods are widely used, for example, for recommendation systems, network analysis, and anomaly detection [20].

- **Reinforcement learning** (RL) learns from rewards and penalties to optimize actions that meet a specified goal. It is used in situations where there is no single "correct" output or action, but there are multiple "good" outputs [19]. The advantages of RL are its ability to solve complex real-world problems and accurate results. However, RL is not preferred for simple problems, the algorithm requires a lot of data and computations, and if the model is too reinforced, it can lead to an overload of states which can weaken the results [20]. Common algorithms are Q-learning and proximal policy optimization (PPO) [19].

  Common applications of RL include video games, resource management, and robotics [20].

- **Semi-supervised learning** tackles problems in which only a small portion of the data is labeled. It combines these labeled examples with a larger pool of unlabeled samples to improve performance, especially in scenarios where labeling is resource-intensive, for example costly or time-consuming. Semi-supervised learning algorithms are generally categorized into three groups: transduction, induction and inherently self-supervised [19].

Figure 2 visualizes the main types of machine learning described above. This is not an exhaustive list, and there are many other ways to categorize ML methods. Some sources focus on the first three categories (from left to right), while others include four or more types. In summary, supervised learning is applied when labeled data is available, whereas unsupervised learning is used with unlabeled data. Reinforcement learning is suitable for problems that can be framed in terms of rewards and penalties, and semi-supervised learning is used when the data includes both labeled and unlabeled examples.

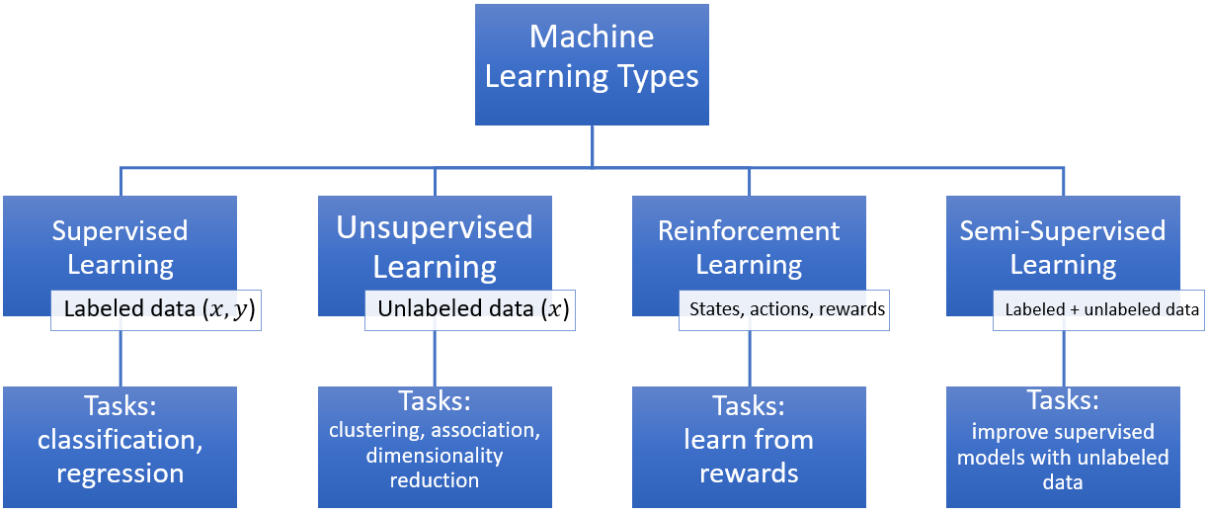


Figure 2. Main types of machine learning methods.

2.3 Machine Learning Algorithms

This section explains the specific ML algorithms used in this thesis. Many different ML methods exist beyond these.

2.3.1 Decision Trees

Decision trees (DTs) are considered transparent (white box) models. They do not necessarily need XAI methods (e.g., LIME or SHAP) to explain their results due to the data being visually understandable, assuming that the tree is not highly complex [22]. DTs are divided into classification trees for yes/no (Boolean) outcomes and regression trees for numerical predictions [23].

Each tree has a root node, branches, internal nodes, and leaf nodes, which makes the model easy to interpret. At each internal node, the model splits the data using a Boolean decision

rule. This process continues until it reaches a leaf node, which displays the final outcome. To find the optimal split, the model uses criteria such as Gini Impurity and Entropy. Gini Impurity indicates how pure or impure a node is with respect to the class labels it contains; a lower value means the features have split into smaller and more precise categories. Entropy measures the amount of disorder and uncertainty, and the tree minimizes this by choosing features that provide the most information [23].

The primary advantages of DTs are their simplicity and transparency as white-box models. They require relatively little data preprocessing and can handle both numerical and categorical variables. However, DTs also have certain drawbacks, such as overly complex trees, instability, poor generalization, and potential biases when some classes dominate the training data [24]. Figure 3 visualizes a simple example of a balanced decision tree. In practice, DTs can become more complex and are rarely perfectly symmetrical.

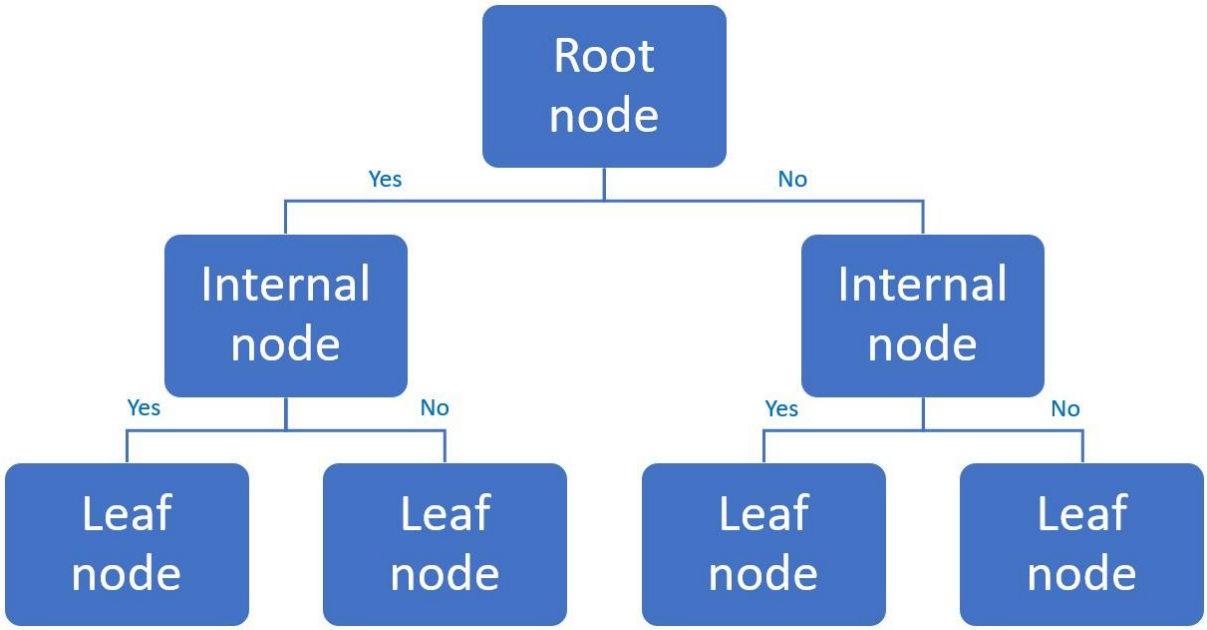


Figure 3. Example of a balanced decision tree model.

2.3.2 Support Vector Machines (SVMs)

Support vector machines (SVMs) are supervised ML methods that can be used for both classification and regression. Their goal is to learn a decision boundary, called a hyperplane, that maximizes the distance between the different classes, relying on support vectors to define

this margin, as shown in Figure 4 [25]. SVMs are commonly divided into two main types: linear SVMs and non-linear SVMs.

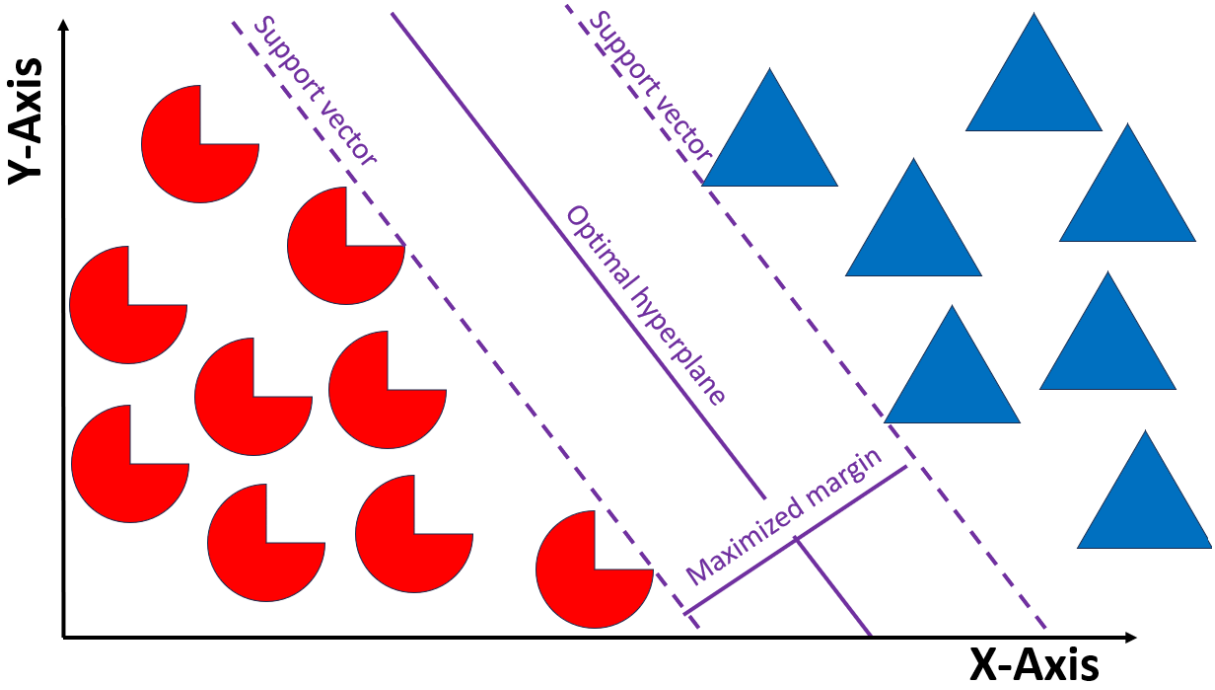


Figure 4. Example of a linear SVMs classification. To separate the two classes as widely as possible, the algorithm positions an optimal decision boundary based on the closest data points, known as support vectors.

Real life applications of SVMs include text classification, image classification, bioinformatics, and geographical information systems [25]. In healthcare, for example, SVMs can be used to predict whether a tumor is benign or malignant. Their advantages include strong performance in high-dimensional spaces, the ability to model nonlinear decision boundaries, robustness to some outlier, support for both binary and multiclass problems, and relatively efficient memory usage. However, SVMs also have drawbacks: training can be slow, parameter tuning is often challenging, models can be sensitive to noise and feature scaling, and the resulting decision function is typically difficult to interpret [26].

2.3.3 K-Nearest Neighbors (KNN)

The k-nearest neighbor (KNN) is a simple and widely used supervised ML algorithm, and it is typically used for classification tasks but can also be applied to regression tasks. It operates by identifying the most frequently represented label among the k-nearest neighbors of a given data point, where the *k* represents the number of nearby neighbors considered when making a

prediction. The advantages of KNN include ease of implementation, it is adaptive and does not require much data to operate. Disadvantages include scaling problems, it does not work well with high-dimensional data inputs, and it is prone to overfitting [27] [28].

Figure 5 illustrates the basic principle of the KNN algorithm, where neighbor count of k = 4. The algorithm identifies the four closest data points (yellow circle) and performs plurality voting (commonly referred to as “majority voting”).

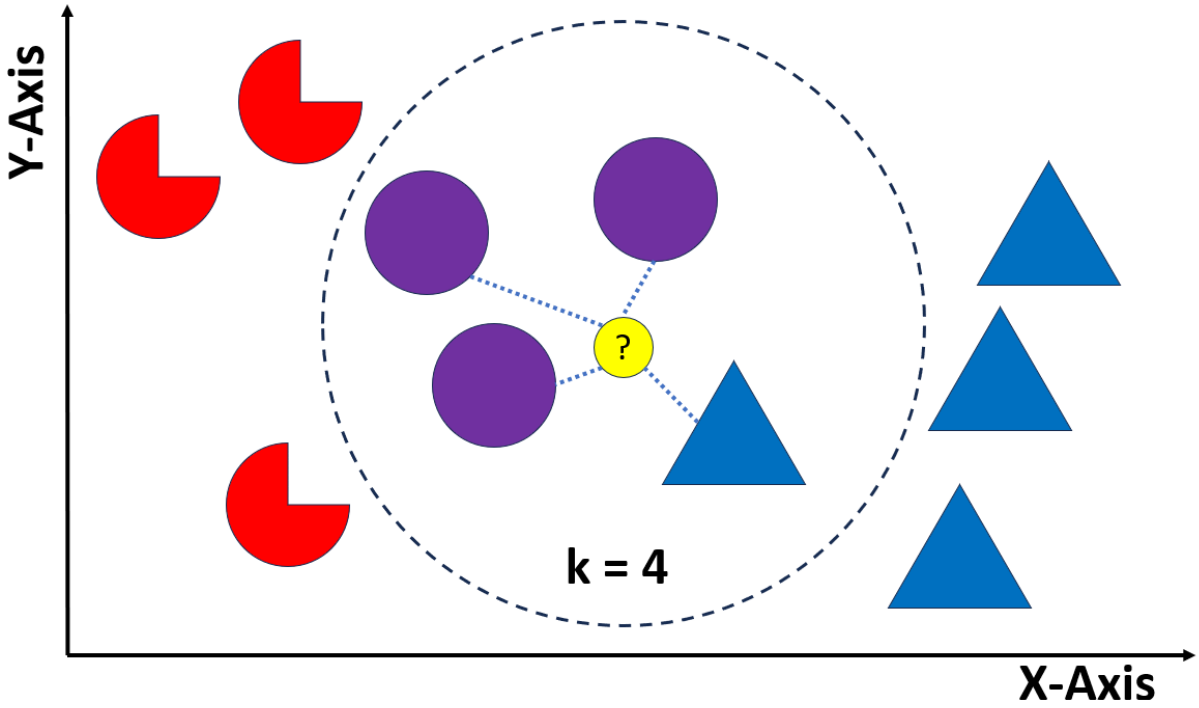


Figure 5. Example of KNN classification using four nearest neighbors (k = 4).

The distance between data points can be calculated using, for example, Euclidean, Manhattan, Hamming, and Minkowski distances. Real-world applications of KNN include, for example, healthcare, data preprocessing, recommendation systems, finance, and pattern recognition [28].

2.4 Deep Learning

In deep learning (DL), models are trained to work with raw data and to automatically identify patterns and features that support relevant to a given task [16]. DL is built on artificial neural networks, which are loosely inspired by the structure and functioning of the human brain. In biological systems, the brain transmits electrical and chemical signals across networks of interconnected neurons. In DL, these “signals” correspond to weighted outputs of

mathematical operations performed by artificial neurons (nodes). A DL model can be viewed as a series of nested mathematical functions that map inputs to outputs. In comparison to traditional ML techniques, DL needs large amounts of training data and substantial computational resources. The negative side of neural network-based approaches is their reduced interpretability as it is often difficult to understand how the model arrived at its predictions. For this reason, DL models are often referred to as "black boxes" [29].

Deep learning methods can be divided into a selected list of different models listed below [29].

- **Convolutional neural networks** (CNNs) are a type of DL architecture tailored for visual data. They are widely used in computer vision tasks, for example for object detection, image recognition, image segmentation, and image classification [29].
- **Recurrent neural networks** (RNNs) are designed for sequential data and are commonly applied to tasks such as speech recognition or neural language processing (NLP) [29]. NLP is used widely in many applications such as powering search engines, chatbots, voice-operated digital assistants, for example, Amazon's Alexa, Apple's Siri, and Microsoft's Cortana [30].
- **Transformer models** are usually associated with large language models (LLMs) [29]. These models are the next evolution step of RNNs. OpenAI's ChatGPT model, GPT-3, started the modern era of generative AI (GenAI). The GPT stands for "Generative Pre-trained Transformer" [31].
- **Diffusion models** are widely used for image generation and can also be adapted to produce other types of content, including text, video and audio data [29]. Notable example of a diffusion model is OpenAI's DALL-E.

Figure 6 visualizes the main types of deep learning described above. This is not an exhaustive list, and there are many other ways to categorize DL methods. In summary, CNNs are used when computer's vision is needed, RNNs are used for NLP applications, transformer models are used for LLMs, and diffusion models are used for image generation.

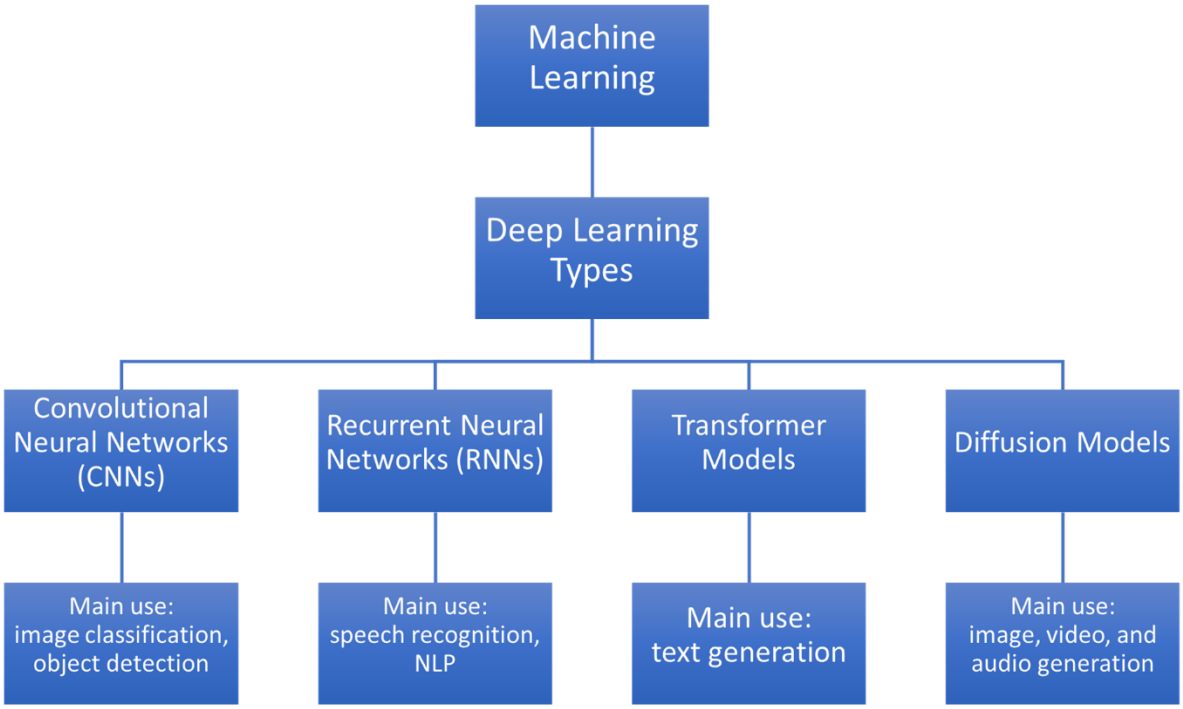


Figure 6. Main types of deep learning methods.

2.4.1 Large Language Models

Large language models (LLMs) are a subcategory of deep learning models based on transformer models. They can interpret natural language and other forms of input to carry out many different tasks, including generating text, summarizing information, and answering questions. Transformer models are a neural network architecture that plays a key role in how LLMs operate. LLMs are a type of generative artificial intelligence (GenAI), and they can produce novel content. One of their key strengths is the ability to respond to open-ended questions without being explicitly programmed for the tasks [32].

LLMs predict the probability of the next word in a sequence repeatedly and learn patterns from text when generating answers. Common use cases for LLMs are, for example, text generation, text summarization, code generation, and language translation. LLMs are trained on vast amounts of data – billions or even trillions of words – from various text sources like online articles, books, websites, and code. They are trained with self-supervised learning and later they are fine-tuned. Retrieval-augmented generation (RAG) allows a pretrained LLM to link external knowledge sources so that it can produce more relevant and accurate responses [32].

While LLMs are powerful tools, they have several limitations. LLMs can hallucinate, generating false or misleading information while sounding convincing and plausible. Biases are another concern, as LLMs can reflect and amplify them, generating unfair or offensive outputs that are presented in LLMs' training data [32]. Additionally, LLMs are vulnerable to data poisoning where the training data is intentionally manipulated to introduce vulnerabilities, biases, or backdoors. Common examples of such vulnerabilities are malicious data injection by adversaries, users unknowingly injecting harmful content, and unverified training data [33].

2.5 Artificial Intelligence in the Healthcare Sector

AI applications are often integrated into clinical workflows, where they assist healthcare professionals rather than replace them. ML algorithms are applied to improve healthcare data accuracy and efficiency. Patient records contain a lot of unstructured data that would take much effort to evaluate, but with the support of ML, healthcare professionals can more easily predict health issues [34]. It is used for classification, prediction, and clustering tasks over vast amounts of complex datasets. Compared to traditional biostatistical approaches for analyzing data, ML can offer improved performance in a variety of tasks. When ML is combined with mobile health solutions, such as smartphones and wearable devices, it can enhance predictions and network efficiency even more. For example, the use of ML in healthcare sector may assist with predictive analytics, diagnosis, and treatment. ML methods can also help to find better personalized medicines and assist as with clinical decision [35].

Supervised and unsupervised learning approaches have demonstrated significant promise in healthcare applications. Supervised approaches are used to estimate the risk of heart-related diseases, classify medical images, and identify cancerous cells. Meanwhile unsupervised methods are applied to tasks such as anomaly detection and clustering, identifying rare disease patterns, and extracting relevant features from medical images [35].

While previously mentioned methods can be valuable tools in healthcare, it is also important to remember that they are not perfect. For instance, supervised learning approaches requires a vast amount of labeled data to perform reliably. Biased or unrepresentative data may lead to inaccurate or unfair predictions. For unsupervised learning, the result may not always be clinically meaningful, and the outputs may be difficult to interpret [35].

LLMs have also seen increasing use in the healthcare sector as they enable rapid summarizations and rephrasing of information that would otherwise take time and effort of professionals [36]. Another potential application of LLMs is in the field of psychology, where they may improve the understanding of psychological characteristics. They rely on behavioral data, are highly scalable, and allow for more flexible response formats. Such assessments can be conducted through everyday activities, for example via smartphones [37].

2.6 Explainable Artificial Intelligence (XAI)

In this thesis, the terms interpretable machine learning (IML) and explainable AI (XAI) are used to refer to closely related concepts that aim to make model behavior understandable to humans. The key difference between the terms is that IML is used specifically in the context of machine learning, and XAI is used in a broader AI context. Both concepts are closely related and contribute to improving interpretability, reliability, and trustworthiness of AI systems.

Especially in the healthcare sector, it is essential to grasp how AI systems generate predictions, as they impact directly on the patient. Without explainability, it is nearly impossible to backtrack and evaluate the reasoning behind an AI prediction. The goal of XAI is to clarify how AI systems arrive at their decisions, making these processes more transparent and interpretable, while also assisting in uncovering possible biases [38].

There are different approaches to interpreting ML models. Black-box ML models are opaque to the observer, whereas white-box models are transparent. White-box ML models are typically simpler and easier to understand but may have limited capacity to capture complex patterns in large datasets. Black-box models are more capable of handling complex datasets, at the cost of reduced interpretability [38].

XAI can be distinguished into two main types of explainable techniques, interpretability by design and post-hoc interpretability. Some basic interpretabilities by design approaches are linear regression, decision trees, and decision rules. Unlike inherently transparent models, post-hoc XAI operates as a secondary analysis layer once the training phase concludes [39].

2.6.1 Post-Hoc Interpretability

Post-hoc interpretability (or post-hoc explainability) methods are applied after an ML model has been trained with the training data by analyzing how different input features influence the

model's predictions. These methods are particularly useful for complex black-box models. Common post-hoc methods include feature importance explanations and counterfactual explanations [40]. Post-hoc methods are often grouped into two categories: model-agnostic and model-specific. In the model-agnostic case, the method ignores internal details of model and explores its behavior systematically changing input values and monitoring the resulting outputs, such as through feature permutation. Model-specific approaches are designed for a specific type of model. For example, they can be used to analyze which types of inputs activate a neural network [39].

LIME and SHAP are examples of feature importance explanations that provide quantitative values of how individual features contribute to a model's predictions. The purpose of feature importance is to identify which features have the greatest influence on the model's output. Features that cause larger changes in the prediction are considered more influential and help identify the key factors of a prediction [40].

Counterfactual explanations highlight what small changes to the input would have been needed to obtain a different prediction, providing actionable insights. These explanations can support decision-making by offering "what-if" scenarios. However, generating meaningful counterfactual explanations can be difficult and extra attention is needed to ensure realistic results [40].

Visualization can be pivotal when interpreting the results and making the results into more accessible and human-friendly forms. When combined with XAI methods, they can further improve the trustworthiness and transparency of a ML model [40].

2.7 XAI Methods

2.7.1 Local Interpretable Model-agnostic Explanations (LIME)

The first post-hoc method used in this thesis, local interpretable model-agnostic explanations (LIME), was proposed by Ribeiro, Singh, and Guestrin in 2016 to explain the predictions of ML models [41]. It can be applied to tabular data (data is in tables), text data, and image data. LIME operates by slightly modifying the input data and observing how these changes influence the model's output. LIME explains an individual prediction by fitting a simple, interpretable model locally around the instance to mimic the behavior of the more complex model in that region of the feature space [42]. Using this information, LIME generates a new

interpretable model that estimates the behavior of the original model, thus providing insight of model's decision-making [39].

The mathematical formula of LIME explanations can be expressed in different ways depending on the source and implementation, but the underlying principle remains the same. Below is one way to present it [43].

2.7.2 Shapley Additive Explanations (SHAP)

Another post-hoc method, Shapley additive explanations (SHAP), was introduced by Lundberg and Lee in 2017 to explain individual predictions [44]. SHAP builds on cooperative game theory and uses Shapley values to attribute the output of any ML model to its input features. In game theory, outcomes are analyzed in terms of interactions between multiple "players". In the SHAP setting, each feature of an input instance is treated as a player in a prediction game, and the Shapley value of a feature reflects its average marginal contribution to the prediction when it is added to different subsets of the other features [39].

2.7.3 Explain Like I'm 5 (ELI5)

Third post-hoc method used in this thesis is explain like I'm five (ELI5) Python module by Mikhail Korobov and Konstantin Lopuhin [46], which assists with ML predictions in an easily understandable way. It is useful for visualization and debugging various ML models [46]. A major limitation for ELI5 is that it works only for linear or parametric and tree-based models and does not support model-agnostic interpretations [47].

3 Cybersecurity and Privacy

This chapter provides a broader perspective on cybersecurity and privacy and links them to a larger framework as these concepts are not specific to the healthcare domain. Cybersecurity threats, privacy risks, and data protection principles are generally applicable across multiple sectors, both within and outside of the EU. This chapter first introduces general cybersecurity considerations, followed by relevant EU regulatory frameworks, and then connects these principles to the healthcare sector. Familiarity with privacy frameworks is necessary as they form the basis for understanding technical vulnerabilities and legal requirements related to data processing.

3.1 Cybersecurity Considerations

Cybersecurity and privacy are essential in today's world and play a crucial role in how we build our systems. Yet, building secure systems is not easy and even naïve mistakes can happen. Recently, the EU's official digital identity wallet prototype could be compromised in under two minutes, revealing significant privacy and security weaknesses in its implementation. The application is planned to be used for age verification across the EU. A user attempting to bypass the age verification could simply modify the values in the configuration file with minimal technical skills [48]. Considering that the EU has some of the strictest security and privacy standards in the world, it is bizarre that such a design flaw can still occur in an application developed at the EU level. This example shows that even systems developed under strict regulatory frameworks may contain security weaknesses. Security and privacy should not be taken for granted but instead require careful design and implementation.

Cyberattacks against healthcare sector are a direct threat to patient safety by causing medical devices such as insulin pumps and heart monitors to malfunction and prevent access to electronical medical records. Common methods of attacking include malware, phishing, ransomware, zero-day exploits, brute force attacks, and denial-of-service (DoS) [49]. AI may improve cybersecurity, but it can also weaken security due to AI enhanced cyberattacks.

In the context of this thesis, cybersecurity and privacy are inspected from machine learning and deep learning perspective.

3.2 European Union Legislative Framework

The legislative framework discussed in this chapter focuses primarily on regulations. Regulations are legally binding acts that are directly applied within the EU Member States. They also apply to non-EU companies operating within the EU or processing data of EU citizens.

Other types of EU legislative acts include directives, decisions, and recommendations. Directives are legislative acts that the Member States must implement by adapting their national laws. A decision is binding on specific entities, such as individual Member States or organizations. Recommendations are not legally binding and serve as guidance without legal obligation [50].

3.2.1 The General Data Protection Regulation (GDPR)

Different regions around the globe have different laws, regulations and standards on how personal data should be handled. The General Data Protection Regulation (GDPR) is the EU's answer to protect individuals when their data is being processed by the private sector or most of the public sector. The regulation also applies to organizations outside of the EU, if they are processing EU citizen's personal data. In general, the GDPR gives individuals more control over their personal data and rights. Key points of the regulation are to make access to an individual's own data easier, right to data portability, right to erasure (the right to be forgotten), and right to know when their personal data has been breached. The regulation came into effect on 25 May 2018 [51].

GDPR is often described as the gold standard of data protection and it is considered one of the strictest privacy and security regulation in the world [52]. The maximum penalty for breaking the regulation is €20 million or 4% of a company's global annual revenue, whichever is higher [53]. In May 2023, social media platform provider Meta (formerly known as Facebook) was fined a record €1.2 billion for failing to comply with the GDPR [54]. In comparison, United States (US), California's version of data protection, the California Consumer Privacy Act (CCPA) fined The Walt Disney Company for $2.75 million in February 2026 [55]. While these examples are not specific to the healthcare sector, the magnitude of the fines highlights the importance of complying with global data protection laws and regulations.

Table 1 presents selected examples of different data protection laws and regulations around the globe. The GDPR applies to all EU member countries, while, for example, the US does not have a single comprehensive data protection law. The GDPR applies to all personal data, while Health Insurance Portability and Accountability Act (HIPAA) affects only to protected health information (PHI) [56]. Overall, the EU has much stronger and comprehensive regulation compared to US.

Table 1. Selected examples of global data protection laws and regulations, ordered by effective date [56].

Abbreviation	Full name	Effective date	Region
HIPAA	Health Insurance Portability and Accountability Act	August 21, 1996	United States
PIPEDA	Personal Information Protection and Electronic Documents Act	April 13, 2000	Canada
APPI	Act on the Protection of Personal Information	April 1, 2005	Japan
PIPA	Personal Information Protection Act	September 30, 2011	South Korea
GDPR	General Data Protection Regulation	May 25, 2018	European Union
DPA (Kenya)	Data Protection Act (Kenya)	November 25, 2019	Kenya
CCPA	California Consumer Privacy Act	January 1, 2020	California, United States
LGPD	Lei Geral de Proteção de Dados Pessoais	September 18, 2020	Brazil

The GDPR defines personal data as “any information that relates to an identified or identifiable natural person”. Organizations operating under the EU jurisdiction are required to comply with the GDPR when processing such data. A similar term, personally identifiable information (PII), is commonly used in the US, and it is not used in the GDPR context. For example, PII considers medical, educational, financial, and employment information as sensitive data; the GDPR considers any information that can be used for identification, such as IP address, cookie ID, medical history, religion, ethnicity etc. While PII and GDPR’s personal data share similarities, the GDPR definition of personal data covers a much wider range of data than PII. According to the GDPR, data that has been irreversibly anonymized is no longer considered personal data [57] [58] [59].

However, there are limitations and exceptions where the GDPR cannot be fully applied. For example, in the healthcare sector, Article 17 of the GDPR, “right to erasure” (also known as “right to be forgotten”), provides individuals the right to request the erasure of their personal

data, but the healthcare legislation requires patient records to be stored for a certain period of time [60]. This creates a legal and practical dilemma, as data protection regulations and healthcare record retention requirements may overlap. There is no single clear solution, and the issue remains to be seen how the EU solves this issue in the future.

3.2.2 The Network and Information Security Directive (NIS2)

The Network and Information Security Directive (NIS2) came into effect on January 16, 2023, and replaced the previous NIS. The primary objective of NIS2 is to enhance the security of network and information systems within the EU. Compared to the GDPR, which is a data protection directive and affects every sector within the EU, NIS2 is a cybersecurity directive that applies specifically to critical infrastructure and essential services [61]. The sectors covered by NIS2 are categorized into two groups: highly critical sectors and critical sectors. The healthcare sector falls under the highly critical sector [62]. On January 20, 2026, new amendments to the NIS2 were proposed to simplify compliance with EU cybersecurity regulations and risk-management requirements [63].

The European Cyber Security Organisation (ECSO) reports that by 2026, only 21 out of 27 EU Member States have transposed the NIS2 Directive into national law [64]. Penalties are similar to GDPR, reaching up to €10 million or 2% of the entity's total worldwide annual turnover in the preceding financial year, whichever is higher [65].

3.2.3 The Artificial Intelligence Act (AIA)

The European AI Office is center of all AI related technologies across the EU and of its main goals is to strengthen the development of trustworthy AI [66]. In 2024, the EU adopted the Artificial Intelligence Act (AIA) regulation, which sets even more precise rules on how to use AI. The goal of AIA is to ensure trustworthy AI in the EU and mitigate the risks to avoid undesirable outcomes. One of the key points of AIA regulation is that cybersecurity-by-design is legally mandated and is not allowed to be used in the EU if AI system's safety is not sufficient [67]

The AIA recognizes four risk levels for various AI systems: unacceptable risk, high risk, transparency risk, and minimal or no risk. Unacceptable risk AI systems are prohibited under the regulation. These include systems that pose a threat to safety, livelihoods, and fundamental rights. High-risk AI systems refer to systems that may pose a serious risk to

health, safety, or fundamental rights. As of the time of writing (May 2026), these systems are subject to strict regulatory obligations, which will come into effect in two parts, in August 2026 and August 2027. Limited risk AI systems refer to transparency risks, such as the need to inform users when they are interacting with AI-generated content or chatbots. These requirements will also come into effect in August 2026. The AIA does not set rules for minimal risk AI systems. Table 2 provides a summary of the four AIA risk levels and selected examples [68].

Table 2. The four levels of AIA's risk based approach with selected examples [68].

Risk category	**Examples**
Unacceptable risk	• Manipulative or deceptive practices causing harm • Social scoring systems • Biometric categorization of individuals (e.g., ethnicity, religion, political opinions) • When law enforcement uses live identification systems in public spaces
High risk	• Devices whose critical safety functions rely on AI (e.g., AI-based systems used in surgery) • AI tools that analyze biometric information to identify individuals, assign them to categories, or estimate their emotions (e.g., systems used to retroactively identify suspects such as thieves)
Limited risk	• Users must be informed when they are not talking to a human but to an AI system (e.g., chatbots) • Content produced by generative AI must be clearly and visibly identifiable
Minimal risk	• AI-enabled video games • Spam filtering systems

3.2.4 Other EU Cybersecurity Policies

European Union Agency for Cybersecurity (ENISA) is one of the central actors in the EU's cybersecurity field. It was founded in 2004 and aims to enhance the trustworthiness of information and communication technology (ICT) products and contribute to EU cyber policies [69] such as the Cybersecurity Act (CSA). The CSA was adopted in 2019, and it gave ENISA a permanent mandate to strengthen cooperation and crisis management within the EU [63].

The CSA also introduced the European Cybersecurity Certification Framework (ECCF), which defines schemes for certifying the cybersecurity of ICT products, services, and processes. Other EU policies are for example the Cyber Resilience Act (CRA), the Cyber

Solidarity Act, and the Cyber Blueprint. The EU also emphasizes other policy guidance such as secure 5G networks, securing the electoral process, cybersecurity of hospitals and healthcare providers, cybersecurity skills, and awareness [63].

In addition, there are various collaborative cyber communities and organizations that support information sharing, incident response, and policy development. These communities play an important role in strengthening cybersecurity practices across sectors, including the healthcare sector. Table 3 presents a selected list of cybersecurity communities within the EU [63].

Table 3. Various cybersecurity orientated communities within the EU in alphabetical order [63].

Abbreviation	Full name	Purpose
CSIRTs or CERTs	Computer Security Incident Response Teams or Computer Emergency Response Teams	EU Member States are required to maintain operational CSIRTs that cooperate at EU level. Their tasks include monitoring and handling incidents, risk evaluation, and contribution to the CSIRTs network.
ECSO	The European Cybersecurity Organisation	It contributes to building cybersecurity communities and strengthening industrial cooperation across Europe.
ENISA	European Union Agency for Cybersecurity	EU cybersecurity agency providing support, coordination, and contributing to frameworks such as NIS2.
ISACs	Information Sharing and Analysis Centres	It contributes to cooperation and information sharing between cybersecurity stakeholders across different economic sectors.
JRC	Joint Research Center	Contributes to cybersecurity research and taxonomy development at EU level.
Women4Cyber	Women4Cyber	A registry aimed at increasing diversity by connecting women professionals in cybersecurity.
Cyber Dialogues	Cyber Dialogues	Platform for cybersecurity discussions with international EU partners to promote shared policy objectives.

3.3 Cybersecurity and Privacy Threats in Healthcare

Cybersecurity and privacy are strongly regulated in the healthcare sector, yet the inadequate level of cybersecurity has led to compromised healthcare data. Malware such as viruses and ransomware have increased, along with phishing emails, password mismanagement, improper security configurations [70], third-party compromise, and supply chain attacks [71]. Insider threats refers to healthcare professionals who misuse their access privileges to malicious purposes [70]. In Finland, for example, there have been cases where curious healthcare

employees and police officers misuse their access for personal information for snooping purposes [72]. Also in Finland, in 2020, the infamous psychotherapy center Vastaamo data breach happened due to failure to follow the GDPR and the safe processing of personal data [73]. This has led to a situation where some patients have committed suicide after the patient records were stolen and used in extortion attempts [74].

On January 15, 2025, in addition to regulatory frameworks such as the GDPR and NIS2, the EU has introduced a dedicated action plan to strengthen the cybersecurity of hospitals and healthcare providers, as the healthcare sector is one of the most targeted sectors by cyber and ransomware attacks. The action plan is based on four key priorities: enhanced prevention, detection, response and recovery, and deterring of cyber threats. In 2023, total of 309 cybersecurity incidents was reported by Member States. Between years 2021 and 2023, total of 54% of ransomware incidents was targeted for the healthcare sector [75]. There are multiple cybersecurity frameworks and best practices that healthcare organizations can implement, such as Zero Trust and MITRE ATT&CK. Blockchain technology can be used to protect data integrity and create audit trails, as they are transparent and tamper-resilient [71].

Human factors refer to various human elements that influence the effectiveness of security measures. A system is usually referred “only as strong as its weakest link”, and in many cases the weakest link is the human factor. For example, user behavior, decision making, skills, and overall interaction between humans and technology play a key role in cybersecurity [76]. Heavy workloads and constant changes make consistent training challenging and create significant security gaps. To mitigate human errors, constant training is the key solution. Staff should be trained with phishing simulations, password management, and responding to security anomalies [71]. LLMs have made phishing emails ever more realistic.

3.4 Vulnerabilities of Artificial Intelligence

Various AI and XAI methods have vulnerabilities and pitfalls that need to be taken into consideration to mitigate the risks related to confidentiality, integrity, and availability – also known as the CIA triad. AI and XAI related risks include bias, data poisoning attacks, poisoning attack, privacy issues, insider threats, and social engineering [77].

AI bias arises when human biases influence or distort the original training data, which can lead to skewed or harmful model outputs. For example, in healthcare, groups that have a minority representation can skew the AI algorithms resulting in lower accuracy results.

Common types of AI biases are for example: algorithm bias, cognitive bias, stereotyping bias, sample bias, and exclusion bias [78].

Data poisoning attacks refer to a cyberattack where adversaries manipulate or corrupt the ML and DL training data by injecting incorrect or biased data points. In the context of LLMs and GenAI, targeted data poisoning attacks aim to manipulate the outputs in a specific way, for example in an attempt of bypass malware detection. Nontargeted data poisoning aims to weaken the model's ability to process data correctly, for example certain sensors to malfunction [79].

Other data poisoning attacks include label flipping, data injection (which work similarly to SQL injection), backdoor attacks, clean-label attacks, and prompt injections. To counter these types of attacks, data validation and sanitization are crucial. Continuous system monitoring and anomaly detection are also necessary [79].

Explainability methods may also introduce privacy-related risks. Although XAI techniques are designed to improve transparency and trustworthiness, they can unintentionally expose sensitive information from the training data. Membership inference attacks attempt to predict whether a specific individual was included in the training dataset based on model's outputs [80].

3.5 Privacy-Preserving Explainable AI Frameworks

Adding explainability to AI models is not enough from a cybersecurity perspective, as privacy concerns must also be addressed. Trustworthy AI aims to mitigate privacy issues related to the data-driven nature of AI systems. The trustworthiness of an AI system depends on key attributes such as explainability, fairness, privacy, reliability, and robustness [81]. These attributes are also closely related to the core principles of the CIA triad.

In addition to data anonymization techniques, cryptographic methods, such as homomorphic encryption and secure multiparty computation (SMC) have been proposed to secure sensitive data in ML systems [82].

The following sections introduce selected privacy-preserving AI approaches and frameworks. While not all of these approaches are directly implemented in this study, they provide relevant background for understanding privacy-preserving XAI systems.

3.5.1 Explainable Privacy-Preserving Intelligent System for Monitoring (X-PRISM)

In the article “Enhancing Privacy Transparency in Remote Patient Monitoring with Explainable AI” the authors propose an Explainable Privacy-Preserving Intelligent System for Monitoring (X-PRISM). Essentially, it could serve as a future framework for IoT based remote patient systems (RPM), similarly to how LIME and SHAP were proposed [83].

X-PRISM is a framework which focuses on protecting patient privacy and maintaining transparency. The system collects real-time health data from various IoT devices and wearables, using SHAP to provide interpretability while ensuring that sensitive data is handled according to the GDPR. This combination of transparency and security is designed to build better trustworthiness [83]. To achieve these goals, the framework integrates federated learning and differential privacy, which are discussed below.

3.5.2 Federated Learning (FL)

In federated learning (FL), multiple clients collaboratively train a shared model in a decentralized way, while keeping their local, potentially sensitive data on-device. Instead of collecting data into a centralized dataset, each node updates the model locally using its own data, while only model parameters are shared. This approach helps mitigate the risk of exposing sensitive data [84].

In the healthcare sector, one of the key challenges for ML is the fragmented nature of medical data, which is often distributed across multiple devices and institutions. FL has been proposed as a solution, but it has limitations such as dataset performance issues and data leakage. To address these limitations, federated deep learning (FDL) has been proposed as a framework for privacy-preserving, improved explainability, and support medical decision-making [85].

3.5.3 Differential Privacy

Differential privacy is used to protect sensitive information during data processing. It works by adding random noise to the data in order to obscure personal information while still allowing the model to produce useful results. The noise can be added to the data either directly to the dataset or during the training process to ensure that the model does not learn sensitive details about individuals. Differential privacy’s main strength lies in its strong privacy protection, however, the balance between privacy and model accuracy is crucial, as extra noise can negatively impact results and degrade model performance [77].

3.5.4 K-Anonymization

K-anonymity is a privacy-preserving technique that prevents individuals from being reidentified in a dataset by ensuring their records cannot be distinguished from those of at least $k - 1$ other individuals. K-anonymization is not to be confused with the k-nearest neighbors algorithm. Data attributes can be divided into three main categories. First, key attributes or identifiers directly reveal a person's identity, such as their social security number or home address. Second, quasi-identifiers, such as age, gender, and ethnicity, do not directly identify someone but may do so when combined with other quasi-identifiers. Third, sensitive attributes contain private information, such as health records, which must remain confidential [86].

Two widely used techniques for implementing k-anonymization are generalization and suppression. Generalization reduces the precision of quasi-identifiers, making it more difficult to combine them for reidentification [86]. Standard generalization value for k-anonymization is $k = 5$, meaning that each group of indistinguishable values must contain at least five individuals to minimize the risk of exposure [87]. Suppression involves removing or masking certain attributes, or even entire records, that could otherwise compromise anonymity [86].

However, using generalization and suppression techniques makes the results less accurate and may introduce bias [86].

4 Methodology

This chapter presents the methodology of the study. It focuses on the experimental setup in JupyterLab, the implementation of the Scikit-learn library, and the dataset modifications. After the modifications, the selected ML are trained and XAI methods are applied to interpret the results.

4.1 Research Approach

This section studies the privacy implications of XAI methods in the context of healthcare datasets. The study is conducted by applying ML models to healthcare-related datasets and analyzing the explanations generated by post-hoc XAI methods (LIME, SHAP, and ELI5). The objective is to evaluate whether these explanation methods may expose sensitive information or introduce potential privacy risks.

The study follows a quantitative and empirical research approach. It is based on experimental data analysis using a preprocessed dataset and machine learning methods implemented in Python. The implementation uses Scikit-learn library, and the development environment is JupyterLab. The tools used in this study, including their version numbers, and brief descriptions are presented later in Chapter 4. The synthetic datasets do not contain real patient information.

Quantitative methods are used to analyze the structured data and to evaluate model performance. The study is empirical in nature, as it relies on observed data rather than purely theoretical models.

Prior to analysis, the datasets were preprocessed and modified to ensure suitability for the research questions and objectives. In the first experiment, a diabetes risk evaluation related healthcare dataset is analyzed. As the dataset is labeled, supervised machine learning methods are used.

4.2 Dataset Description

The dataset used in this study is a modified version of the Kaggle dataset “Diabetes Health Indicators Dataset”, license CC0: Public Domain [88]. Kaggle provides a wide variety of datasets to choose from with different emphasis on features and metrics. For demonstration

purposes the dataset used in this study is a simplified version of Kaggle dataset with limited number of rows and features.

This study examines whether privacy sensitive information can be exposed from the dataset. The dataset sizes (N) used in research are 1,000 synthetic patients. 30% of the values (n) were synthetically modified, resulting in 300 modifications. The 30% rate was selected to provide a clear contrast between the original and modified values while maintaining dataset usability for model training and explanation. These modifications introduce privacy relevant anomalies intended to simulate personal data when the dataset is not properly sanitized in accordance with the GDPR definition of personal data.

There are two intentionally distinguished patients who stand out from the dataset: patient ID 1 represents the oldest patient in the dataset, a 250-year-old male Elf (henceforth patient A). Patient ID 2 represent another high age, a 124-year-old female Elf (henceforth patient B). Patient ID 89 is a 42-year-old Asian female who is the only pregnant patient with gestational diabetes (henceforth patient C).

Diabetes is a chronic disease in which the human body either does not produce enough insulin or cannot use it effectively. Diabetes can be categorized into type 1 diabetes, type 2 diabetes, pre-diabetes, and gestational diabetes. In 2022, approximately 14% of adults aged 18 years and older were living with diabetes globally [89].

4.2.1 Data Preprocessing

The original “diabetes.csv” dataset [88] was preprocessed and modified by the thesis author. The first 1,000 rows of the original dataset were used for analysis. This led to a situation of unrealistic diabetes representation, as over 60% of the patients were diagnosed with diabetes which introduces class imbalance affecting model learning and explanation stability.

For demonstration purposes, fictional ethnicities – Hobbits, Elves, and Dwarfs from The Lord of The Rings fantasy world, written by J. R. R. Tolkien – were added to the dataset. Some of these fictional ethnicities have very high age of over 250 years old. In the ethnicity column, 150 Hobbits, 100 Elves, and 50 Dwarfs were randomly edited, resulting in a total of 300 modified rows. Rest of the attributes were left intact, meaning that the body measurements of these fictional groups are not “realistic” (e.g., unusually tall heights for Hobbits and Dwarfs).

The dataset contains integer, float, and string values. Since string values are not directly accepted by most ML algorithms, they must be converted into numerical representations either within the dataset itself or in the ML code. In this thesis, the string values were converted in the ML code.

Numerical features were standardized using “StandardScaler” to ensure comparable feature scales, particularly for distance-based models such as SVMs and KNN.

To simplify the analysis and the outputs, several variables from the original Kaggle dataset were excluded, including some socio-economic attributes, lifestyle-related measures and additional clinical indicators.

The following features were included for further analysis (Table 4):

Table 4. Preprocessed dataset from the original dataset [88].

Feature name	**Type**	**Description**	**Values/Range**
patient_id	Integer	Unique patient ID	1 – 1000
age	Integer	Patient's age in years (has unrealistically high age)	18 – 250
gender	String	Patient's gender	"Male", "Female", "Other"
ethnicity	String	Ethnic background. Contains additional fictional ethnicities (Hobbit, Elf, Dwarf)	"Asian", "Black", "Hispanic", "White", "Hobbit", "Elf", "Dwarf"
smoking_status	String	Smoking behavior	"Never", "Former", "Current"
alcohol_consumption_per_week	Float	Alcohol consumed per week	0 – 15
sleep_hours_per_day	Float	Average daily sleep hours	4 – 10
family_history_diabetes	Integer	Family history of diabetes	0 = No, 1 = Yes
bmi	Float	Body Mass Index (kg/m²)	15.2 – 42.5
systolic_bp	Integer	Systolic blood pressure (mmHg)	90 – 165
diastolic_bp	Integer	Diastolic blood pressure (mmHg)	51 – 103
glucose_fasting	Float	Fasting glucose (mg/dL)	66 – 163
glucose_postprandial	Float	Post-meal glucose (mg/dL)	70 – 272
insulin_level	Float	Blood insulin level (μU/mL)	2.00 – 23.62
hba1c	Float	HbA1c, hemoglobin A1c (%)	4.00 – 9.53
diabetes_risk_score	Integer	Risk score (0 – 100)	7.3 – 60.5
diabetes_stage	String	Stage of diabetes	0 = "No diabetes", 1 = "Pre-diabetes", 2 = "Type 1", 3 = "Type 2", 4 = "Gestational"
diagnosed_diabetes	Integer	Diagnosed diabetes	0 = No, 1 = Yes

4.3 Implementation Details

4.3.1 Experimental Setup

All experiments were tested in an isolated Python virtual environment (venv) to ensure reproducibility and to avoid interference with the system-wide Python packages. JupyterLab was used as a development environment, with Mozilla Firefox as the browser. Scikit-learn library version 1.8.0 [90] was used for the ML tasks. The explainability methods (SHAP, LIME, and ELI5) were implemented using separate Python libraries. The ML methods used were decision trees (DTs), support vector machines (SVMs), and k-nearest neighbors (KNN). An 80/20 split was applied to the dataset – 80% for model training and 20% for testing. A random state (42) was used to ensure that the results are reproducible in different runs of the experiment.

The datasets were in comma-separated values (CSV) format. Google Sheets was used to edit the data in order to prevent the separator from changing from a comma to a tab or semicolon, as Microsoft Excel kept altering the values during processing due to localization settings. All datasets were formatted using a comma (,) as a separator.

Table 5 presents the tools and libraries, and their version numbers, used in this study.

Table 5. Used tools and libraries and their version numbers. The table is sorted alphabetically by the "Tool" column.

Usage	**Tool**	**Version**
Machine learning	Decision trees (DTs)	See Scikit-learn
Explainability	ELI5	0.16.0
Development environment	JupyterLab	4.5.5
Machine learning	K-nearest neighbors (KNN)	See Scikit-learn
Explainability	LIME	0.2.0.1
Visualization	Matplotlib	3.10.8
Numerical computing	NumPy	2.4.2
Data processing	Pandas	3.0.2
Programming language	Python	3.13.12
Machine learning library	Scikit-learn	1.8.0
Explainability	SHAP	0.51.0
Machine learning	Support vector machines (SVMs)	See Scikit-learn

4.3.2 Dataset Setup in Scikit-Learn

Two experimental dataset configurations were used in the study. In the scripts, these were referred to as “all_features” and “reduced_features” to distinguish the two instances. The first configuration (“all_features”) contained all columns except “patient_id” and “diagnosed_diabetes”. The target variable was “diagnosed_diabetes”, which the models were trained to predict – thus removed. The column “patient_id” was removed to prevent the model from learning from unrelated, non-medical data.

In the second configuration (“reduced_features”), an additional column, “diabetes_stage”, was removed from the feature set. Diabetes stage (no diabetes, pre-diabetes, type 1, type 2, gestational) is a highly informative feature for the prediction task. The purpose of this configuration was to evaluate how the removal of a crucial feature affects model behavior and explainability results in the context of privacy.

4.3.3 Privacy Implementation

After the dataset has been processed with basic configuration and results were saved, privacy-preserving techniques were added. Random noise with a fixed seed value (42) was applied to ensure that the results are reproducible in different runs of the experiment.

Ethnicity was also categorized into two main groups: “Human Group” (Asian, Black, Hispanic, White), containing a total of 700 patients, and “Fantasy Group” (Hobbit, Elf, Dwarf), containing a total of 300 patients.

Table 6 presents the age distribution before and after applying privacy-preserving techniques. Random noise was sampled from a normal distribution with a mean of 0 and a standard deviation of 5.

Table 6. Distribution of patients across age groups before and after privacy-preserving techniques.

Age group	**Before privacy**	**After privacy**
18–29	90	156
30–59	633	521
60+	277	323

To simplify the analysis, “reduced features” configuration was used, which does not include the “diabetes_stage” feature. Thus, no generalization was applied to it.

4.4 Summary of the Methodology

The original dataset was downloaded from Kaggle, which was modified and reduced in size. The modifications included unrealistically high age (up to 250 years) and introduced fantasy world ethnicities to better understand if and when personal information is leaked.

Scikit-learn was used as the primary library for implementing the selected ML models. These models were analyzed to evaluate their behavior using different explainability techniques (see Table 5, p. 34, for details).

The main objective was to investigate if the generated explanations reveal sensitive or personal information, such as age or ethnicity. Three representative patient cases (patient A, B, and C) were introduced for closer analysis to demonstrate potential privacy risks in XAI.

Two experimental configurations were evaluated: a setup using both “all features” and “reduced features” configurations, and a second setup with privacy-preserving techniques, including feature generalization and noise injection to simulate k-anonymity based data protection. The privacy-preserving techniques were applied only to the “reduced features” configuration. The results from these configurations were compared to evaluate the impact of privacy mechanisms on model performance and explainability.

5 Results

In the following sections the results are first presented without additional XAI methods and then followed by LIME, SHAP, and ELI5 results. The terms “all features” and “reduced features” are defined in Chapter 4. To recap, “all features” include all columns except “patient_id” and “diagnosed_diabetes”; in “reduced features”, the “diabetes_stage” column is also removed.

The privacy-preserving experiments were done using only the “reduced features” configuration in order to simplify the analysis and focus on the effects of privacy transformations.

5.1 Decision Tree (DTs) Results

DTs are interpretable by design due to their tree-based structure and no additional XAI methods are needed. Still, XAI methods can be applied to provide additional insight into feature contributions and model behavior.

Figure 7 presents a DT trained with all features (no limits on the tree depth). The tree consists of 3 internal nodes and 4 leaf nodes. The root node is HbA1c, followed by a diabetes stage internal node.

The model’s accuracy score is 1.000.

Gini impurity values range from 0.000 (best) to 0.4999 (worst).

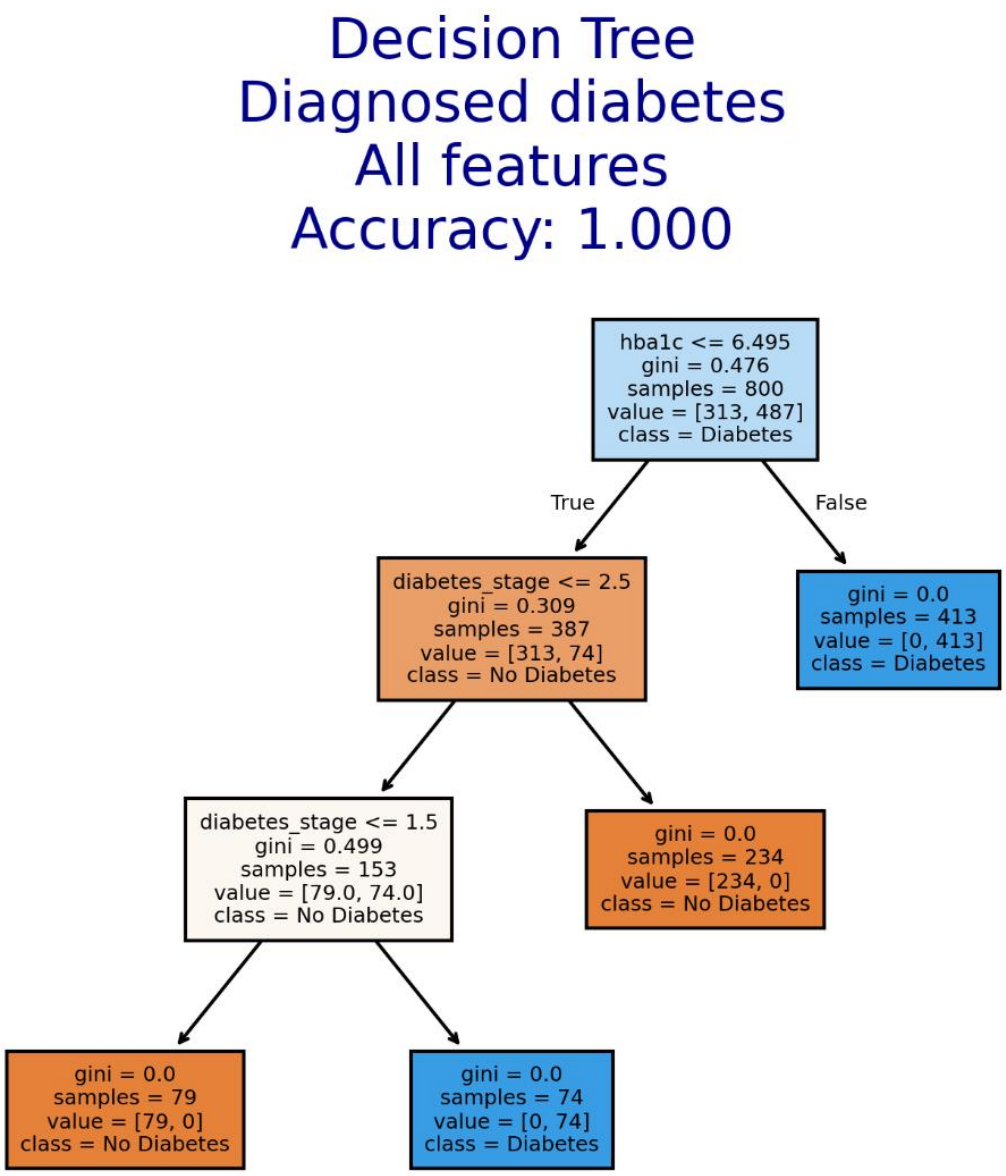


Figure 7. Decision tree with all features.

Figure 8 presents a DT with reduced features and has more internal and leaf nodes. The root node is HbA1c followed by fasting glucose internal node. The tree has maximum depth of 5 and consists of 8 internal nodes and 10 leaf nodes.

The model's accuracy score is 0.910.

Gini impurity values range from 0.00 (best) to 0.499 (worst).

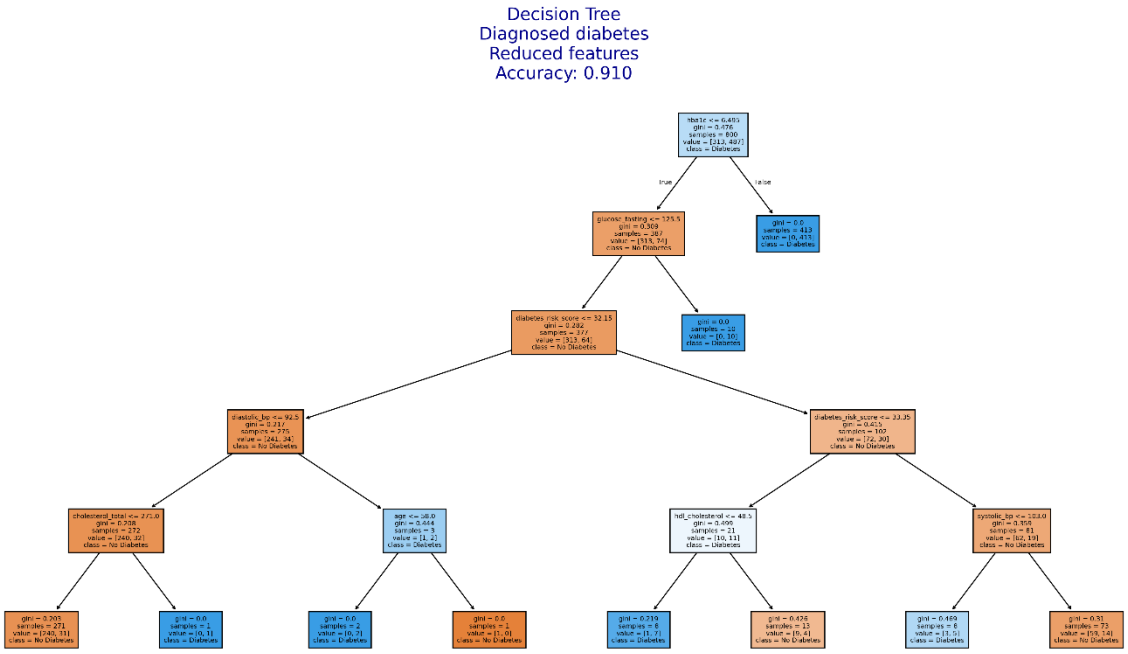


Figure 8. Decision tree with reduced features.

If we take a closer look at the DT (Figure 9), we can see patient counts of 3 or 2, highlighted in yellow. Patient counts of 1 are highlighted in red. These values represent the number of patients in each node.

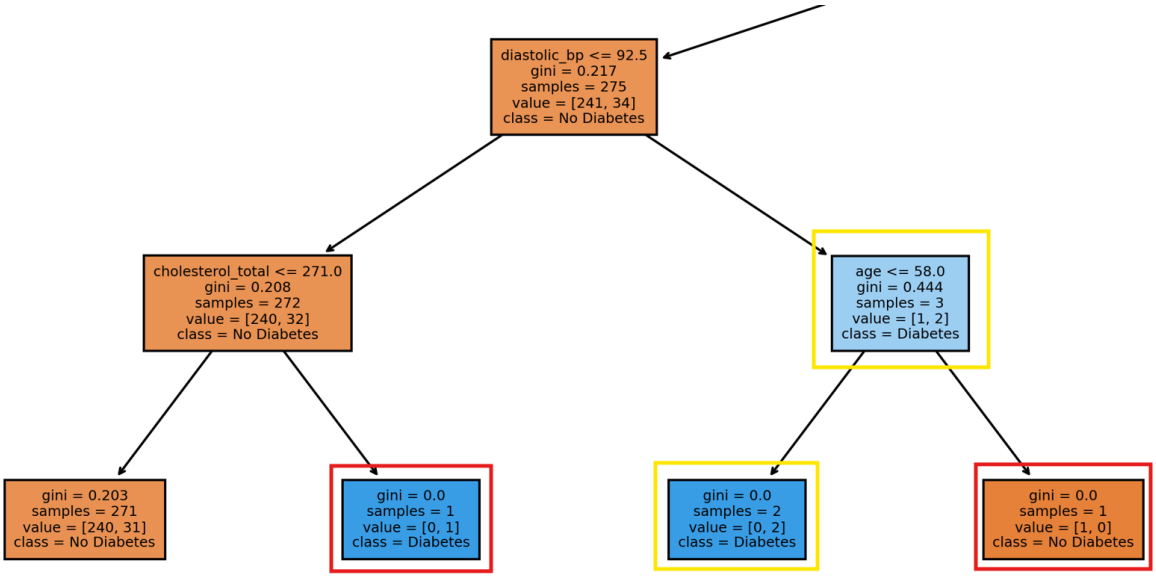


Figure 9. A zoomed picture of the decision tree (reduced features configuration).

Figure 10 presents the DT with privacy-preserving techniques applied. The root node is HbA1c followed by diabetes stage. The tree consists of 2 internal nodes and 4 leaf nodes.

The model's accuracy score is 1.000.

Gini impurity values range from 0.00 (best) to 0.499 (worst).

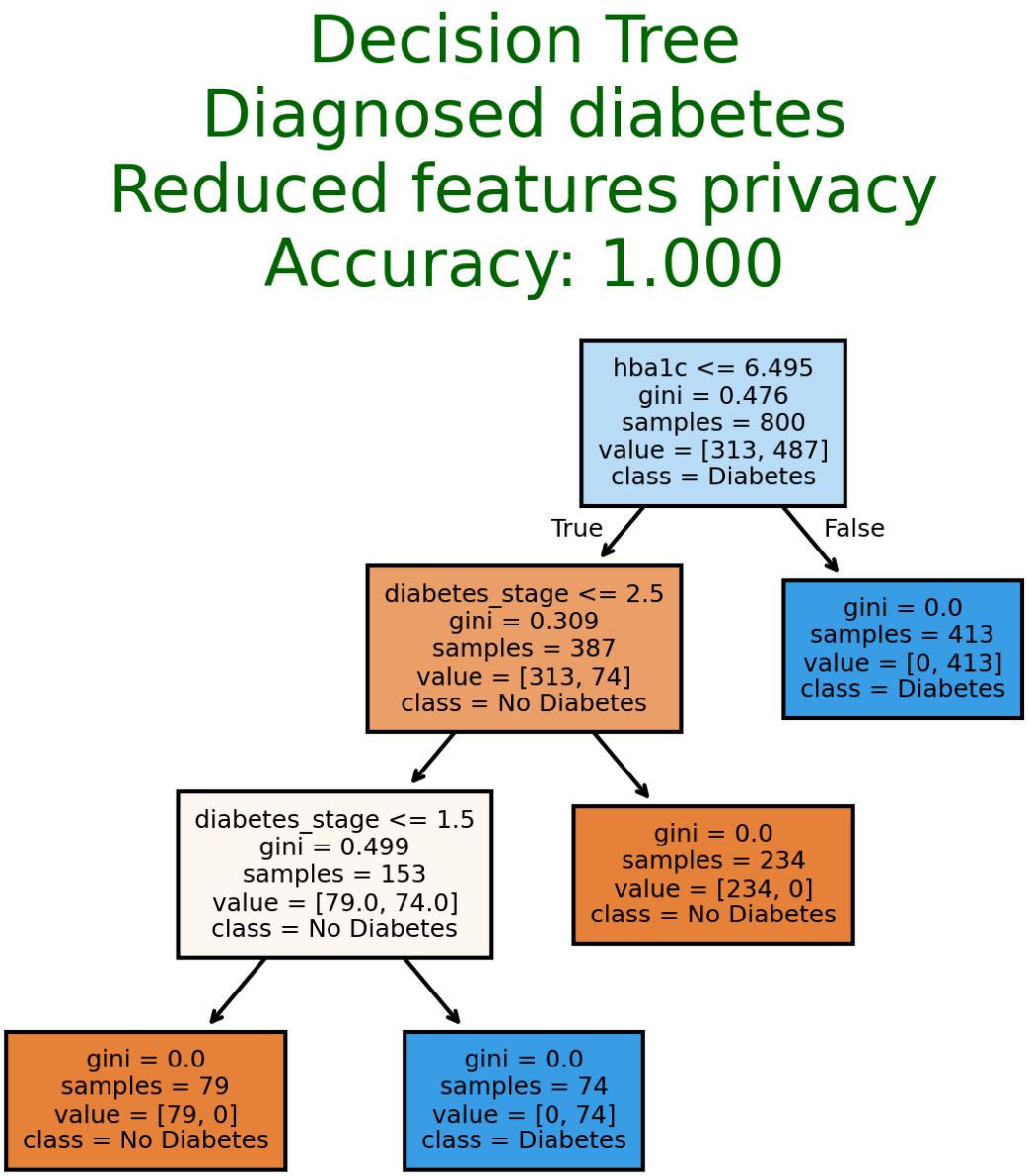


Figure 10. Decision tree with privacy-preserving techniques applied.

5.2 LIME Results

This section presents the results of DTs, SVMs, and KNN models using LIME as the explainability method.

5.2.1 LIME on Decision Trees (DTs)

Figure 11 presents the LIME explanations for the DTs.

Feature names are shown on the left side of the graph and include sensitive information such as ethnicity. The accuracy score of LIME slightly varies between (A) all features and (B) reduced features configurations. For (A) the score is 1.000, and for (B) the score is 0.910.

The model also shows differences in feature contributions.

The top five features for (A) all features are HbA1c, diabetes stage, gender (other), heart rate, and HDL cholesterol.

The top five features for (B) reduced features are HbA1c, fasting glucose, diabetes risk score, HDL cholesterol, and systolic blood pressure.

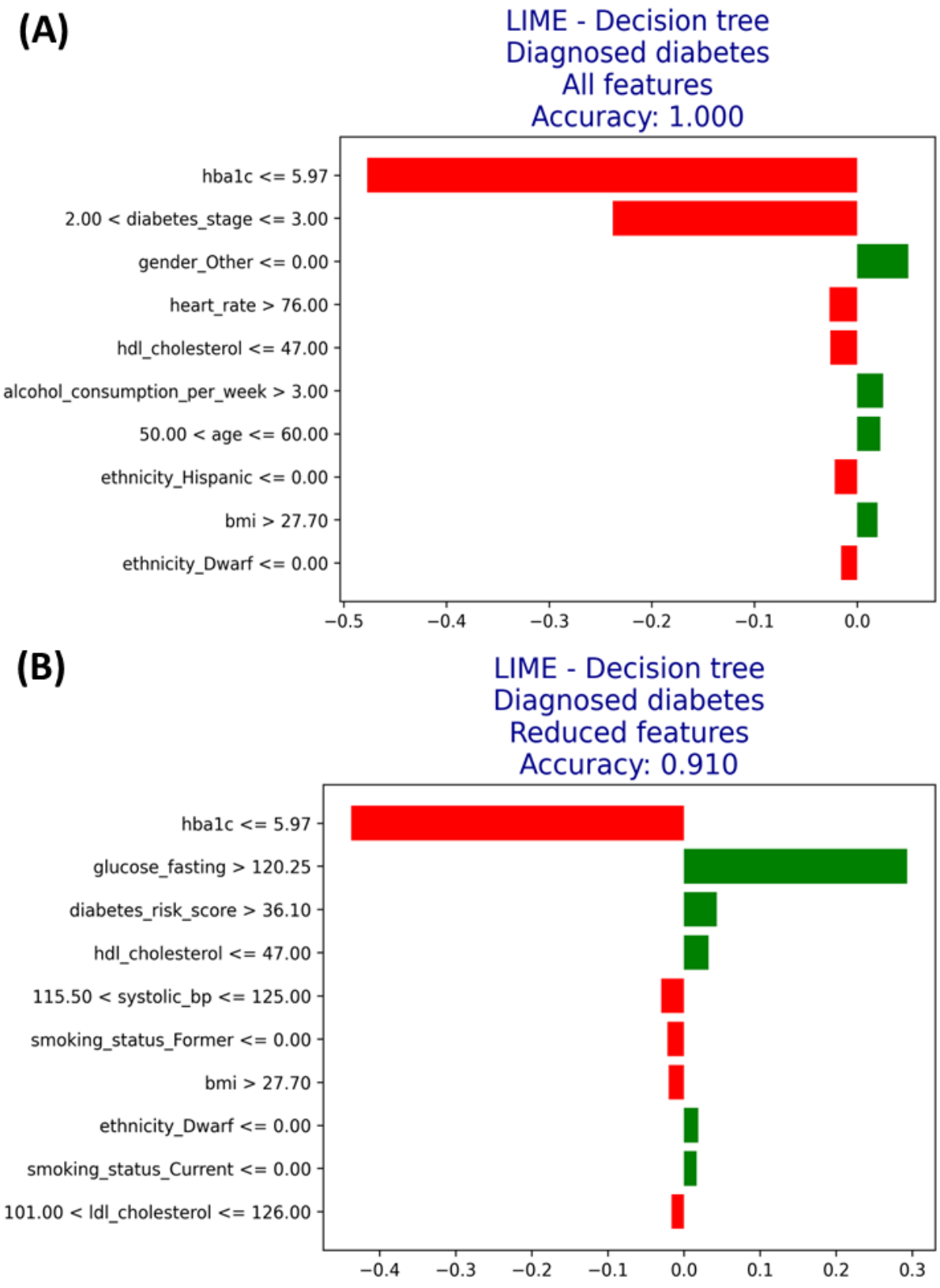


Figure 11. Decision tree LIME feature comparison: (A) full feature set and (B) reduced feature set.

Figure 12 presents the LIME explanations with privacy-preserving techniques applied for the DTs.

Feature names are shown on the left side of the graph and do not contain sensitive information. The accuracy score of LIME is 1.000.

The top five features are HbA1c, diabetes stage, gender (other), smoking status, and gender (male).

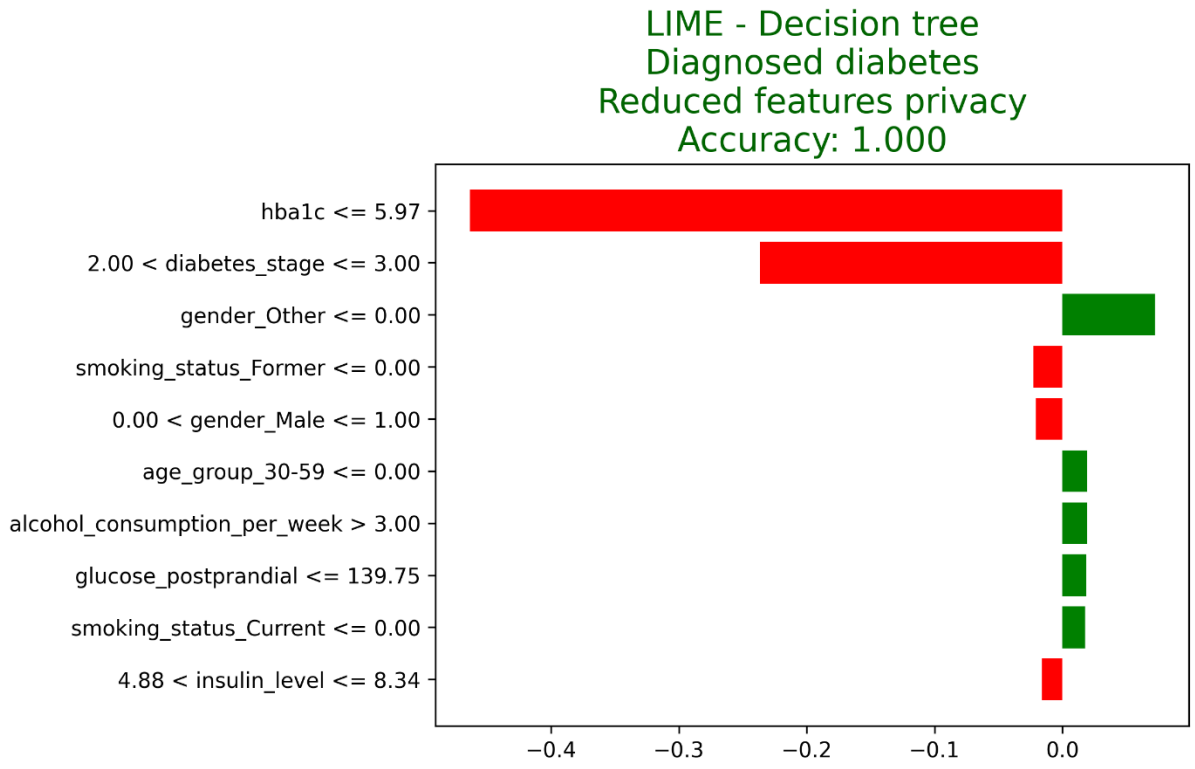


Figure 12. Decision tree LIME with privacy-preserving techniques applied.

5.2.2 LIME on Support Vector Machines (SVMs)

Figure 13 presents the LIME explanations for the SVMs.

Feature names are shown on the left side of the graph and include sensitive information such as ethnicity. The LIME accuracy score for (A) is 0.960 and for (B) is 0.840.

The top five features for (A) are diabetes stage, gender (other), HbA1c, postprandial glucose, and fasting glucose.

The top five features for (B) are HbA1c, postprandial glucose, gender (other), fasting glucose, and diabetes risk score.

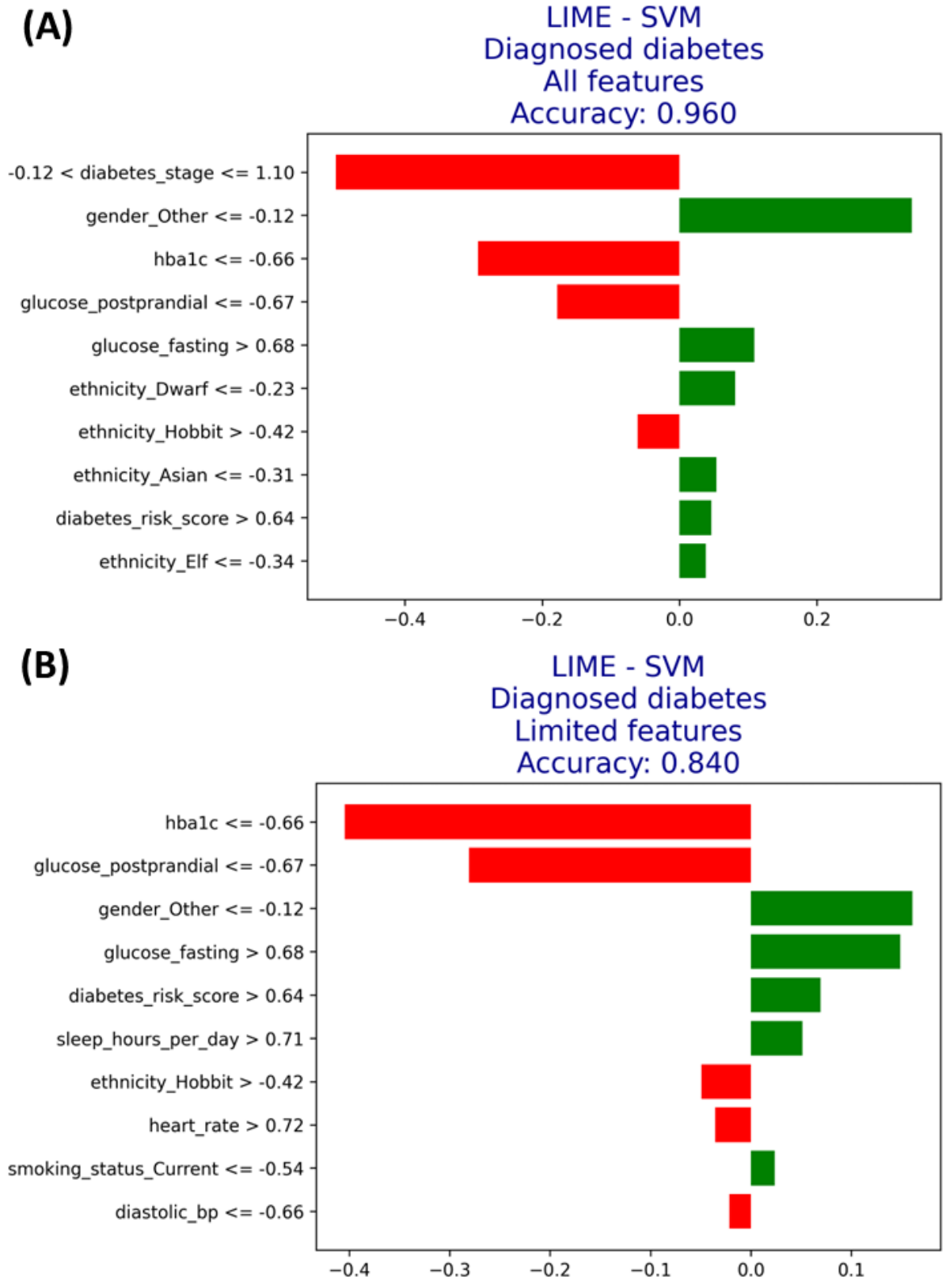


Figure 13. Support vector machines LIME feature comparison: (A) full feature set and (B) reduced feature set.

Figure 14 presents the LIME explanations with privacy-preserving techniques applied for the SVMs.

The LIME accuracy score is 0.855.

The top five features are HbA1c, postprandial glucose, fasting glucose, gender (other), and diabetes risk score.

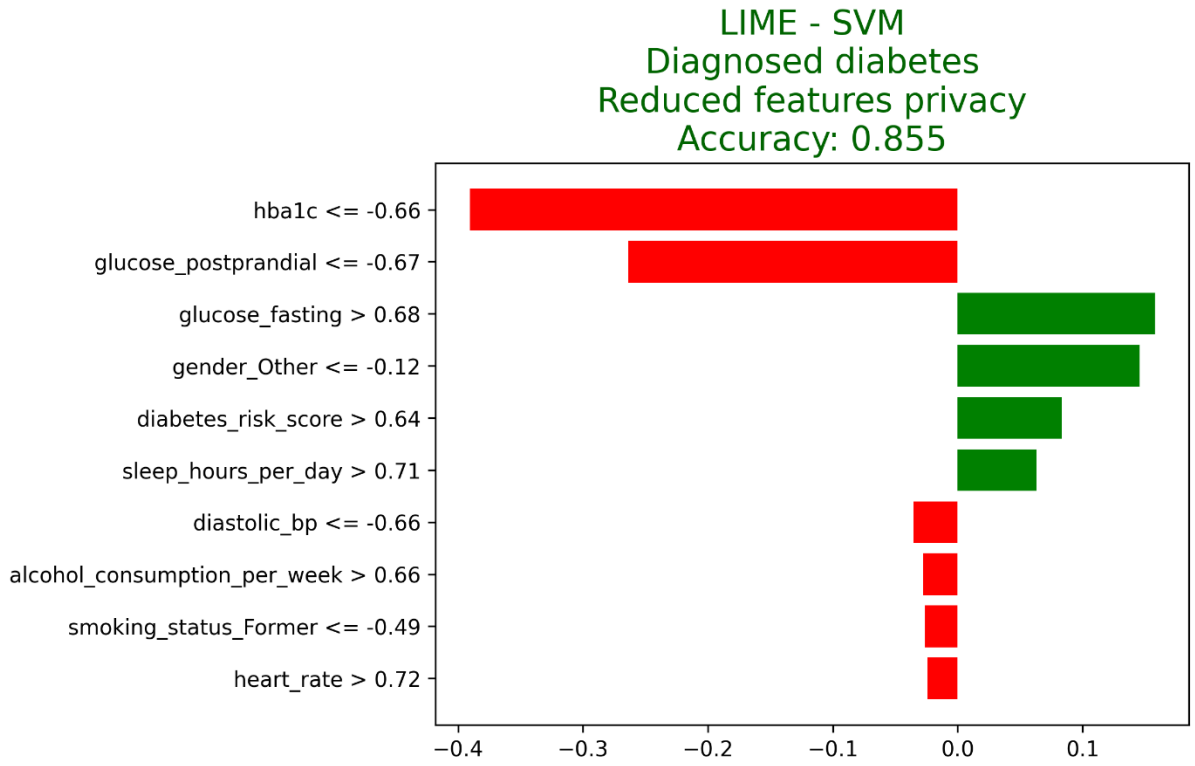


Figure 14. Support vector machines LIME with privacy-preserving techniques applied

5.2.3 LIME on K-Nearest Neighbors (KNN)

Figure 15 presents the LIME explanations for the KNN.

Feature names are shown on the left side of the graph and include sensitive information such as ethnicity. The KNN accuracy score for (A) is 0.785 and for (B) is 0.720.

The top five features for (A) are diabetes stage, HbA1c, gender (other), postprandial glucose, and fasting glucose.

The top five features for (B) are postprandial glucose, gender (other), HbA1c, fasting glucose, and Asian ethnicity.

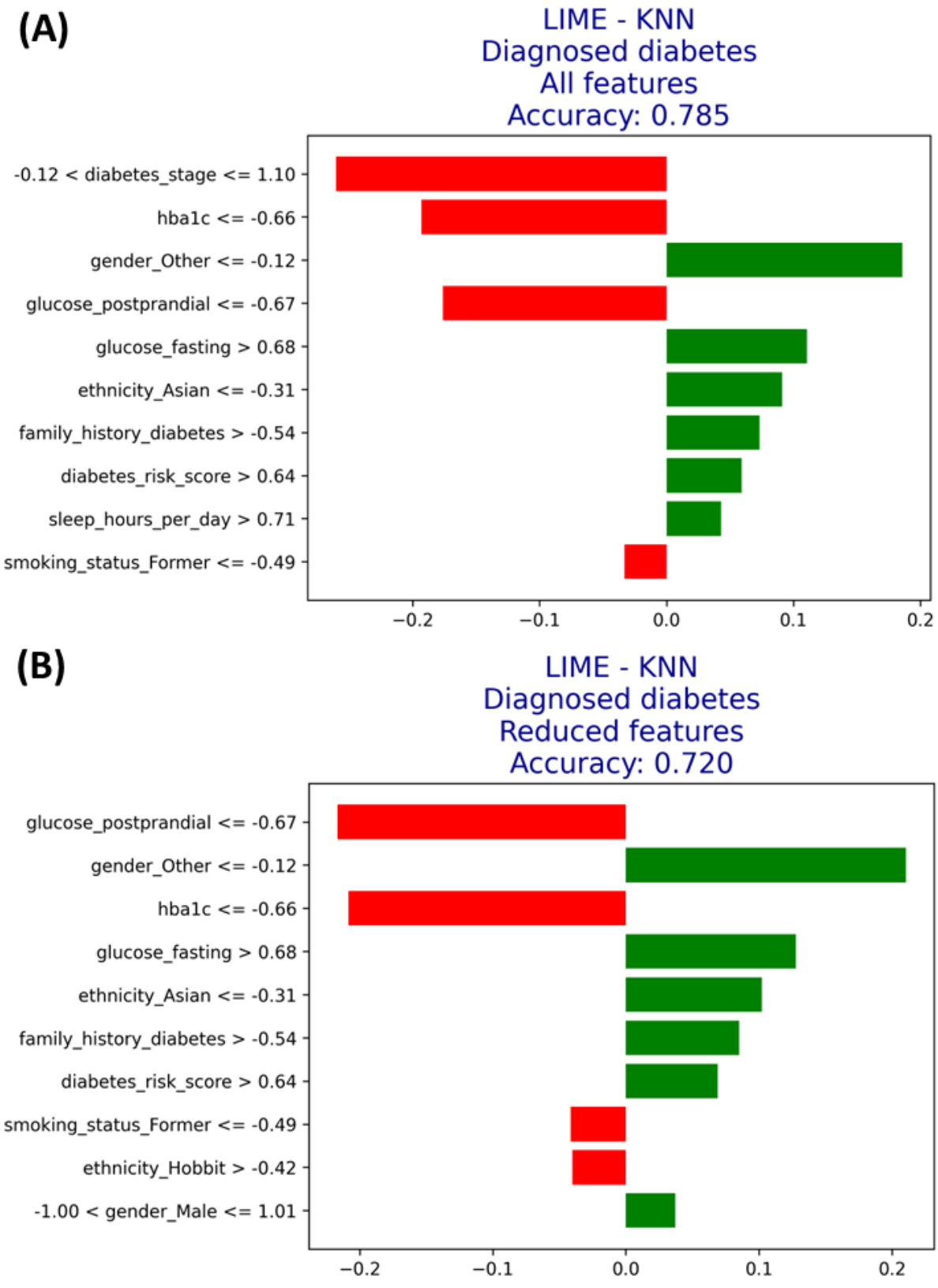


Figure 15. K-nearest neighbors LIME feature comparison: (A) full feature set and (B) reduced feature set.

Figure 16 presents the LIME explanations for the KNN.

The KNN accuracy score is 0.735.

The top five features are HbA1c, postprandial glucose, gender (other), fasting glucose, and family history diabetes.

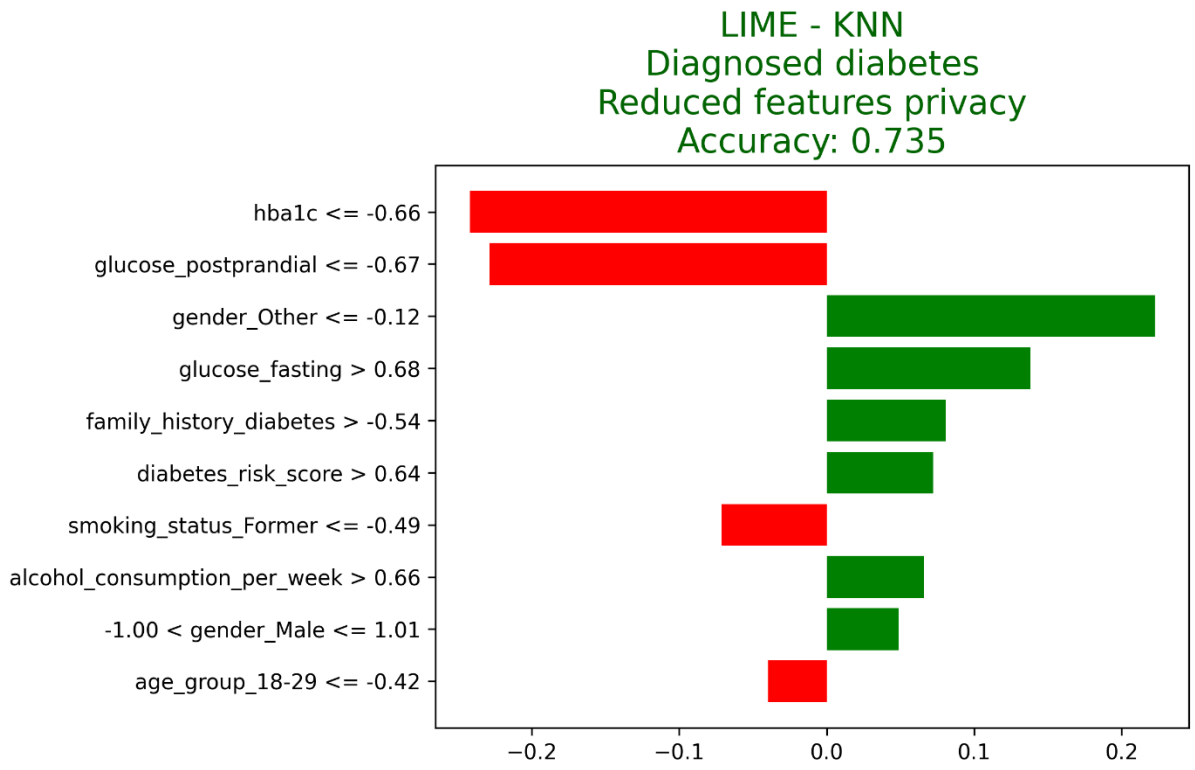


Figure 16. K-nearest neighbors machines LIME with privacy-preserving techniques applied.

5.3 SHAP Results

This section presents the results of DTs, SVMs, and KNN models using SHAP as the explainability method. The explanations include SHAP summary plots and SHAP bar plots.

5.3.1 SHAP on Decision Trees (DTs)

Figure 17 presents SHAP bar plots for (A) and (B). However, due to the limited dataset size, and the characteristics of DTs, the SHAP visualizations do not show the expected variation in the feature importance. Normally, SHAP bar plots show differences in average feature contributions across configurations. In this case, the results appear highly condensed with minimal variation between features. For this reason, only the SHAP bar plot is shown, and the SHAP summary plot is excluded in this decision tree results section.

The SHAP score for (A) is 1.000 and for (B) is 0.910.

SHAP identifies only two features for (A), age and alcohol consumption per week.

SHAP identifies the same two features for (B), age and alcohol consumption per week.

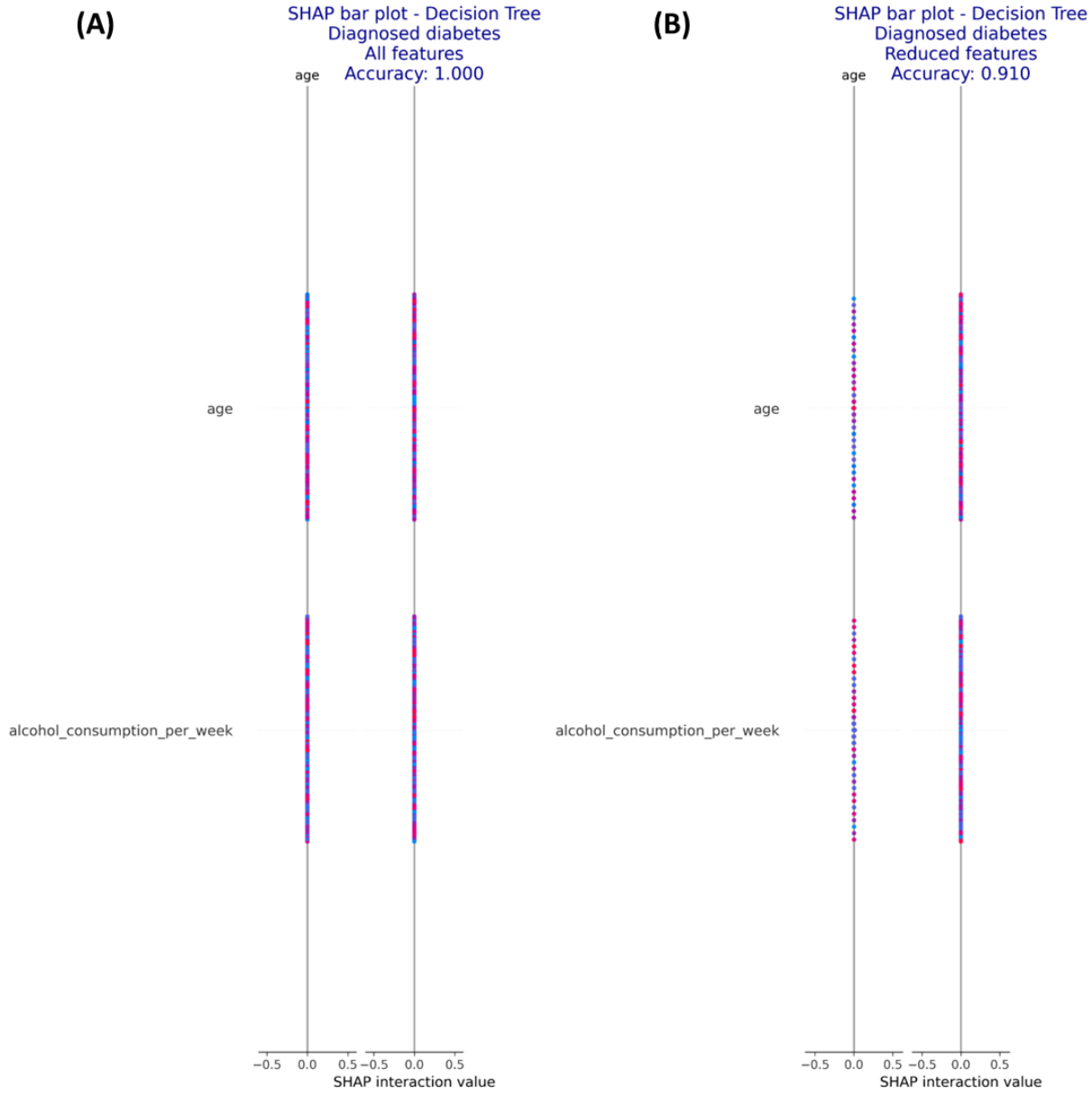


Figure 17. Decision tree SHAP bar plot feature comparison: (A) full feature set and (B) reduced feature set.

5.3.2 SHAP on Support Vector Machines (SVMs)

Figure 18 presents the SHAP summary plot explanations for the SVMs.

Feature names are shown on the left side of the graph and include sensitive information such as ethnicity. The SHAP score for (A) is 0.960 and for (B) is 0.840.

The top five features for (A) are diabetes stage, HbA1c, postprandial glucose, fasting glucose, and Hobbit ethnicity.

The top five features for (B) are HbA1c, postprandial glucose, diabetes risk score, insulin level, and sleep hours per day.

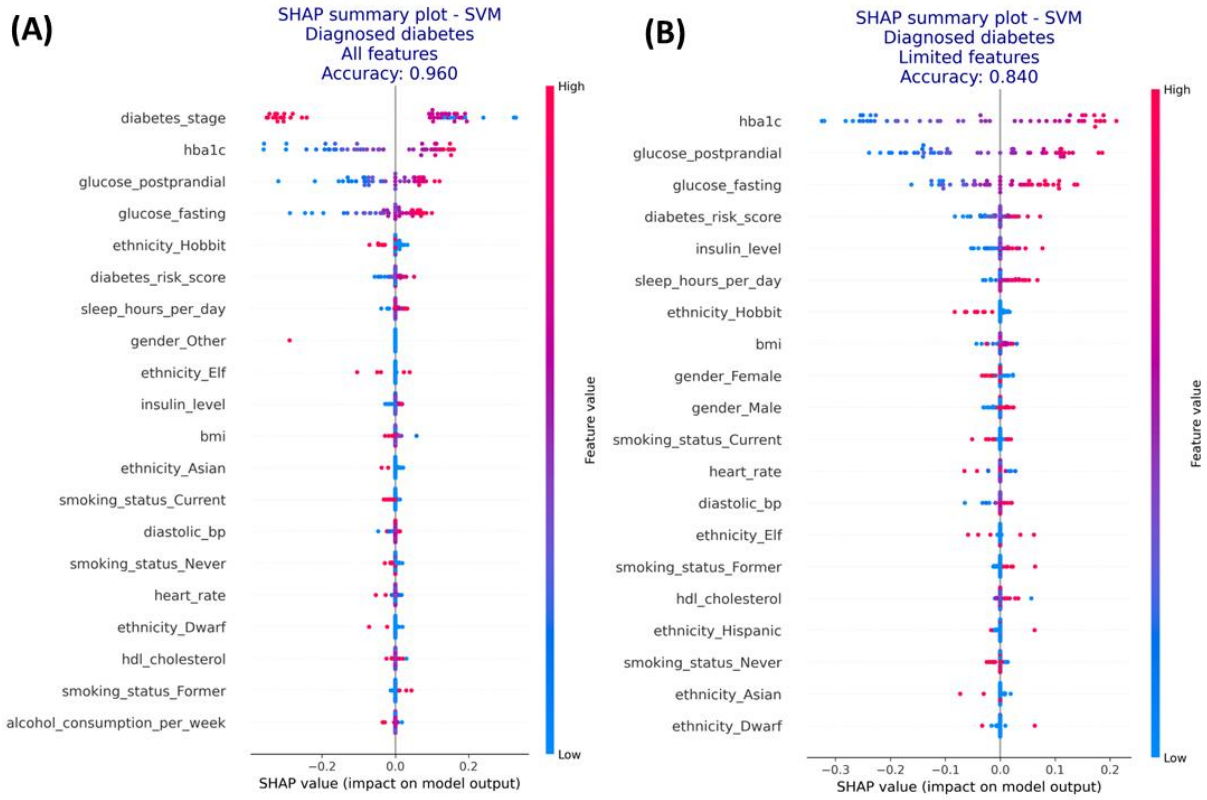


Figure 18. Support vector machines SHAP summary plot feature comparison: (A) full feature set and (B) reduced feature set.

Figure 19 presents the SHAP bar plot explanations for the SVMs. The results are the same as in the SHAP summary plot and only the data presentation is changed.

Feature names are shown on the left side of the graph and include sensitive information such as ethnicity. The SHAP accuracy score for (A) is 0.960 and for (B) is 0.840.

The top five features for (A) are diabetes stage, HbA1c, postprandial glucose, fasting glucose, and Hobbit ethnicity.

The top five features for (B) are HbA1c, postprandial glucose, diabetes risk score, insulin level, and sleep hours per day.

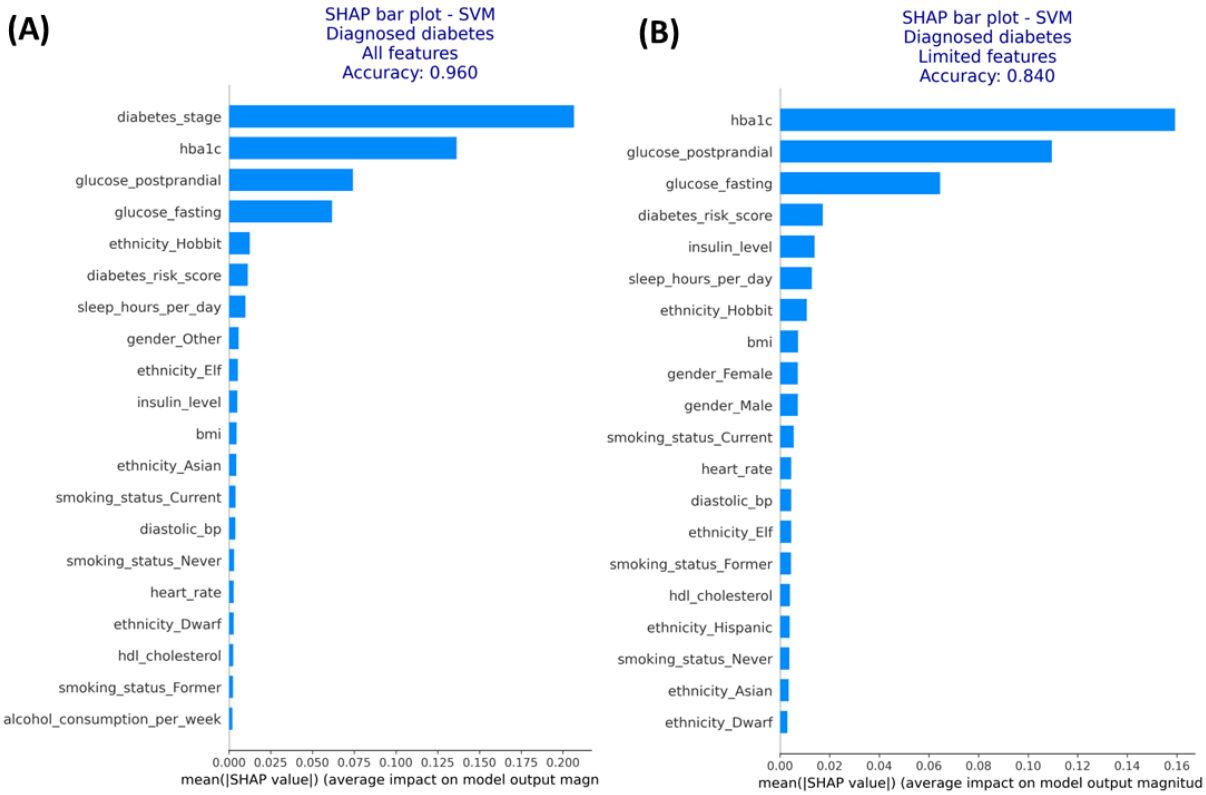


Figure 19. Support vector machines SHAP bar plot feature comparison: (A) full feature set and (B) reduced feature set.

Figure 20 presents the SHAP summary plot explanations with privacy-preserving techniques applied for the SVMs.

The SHAP accuracy score is 0.855.

The top five features are HbA1c, postprandial glucose, fasting glucose, diabetes risk score, and sleep hours per day.

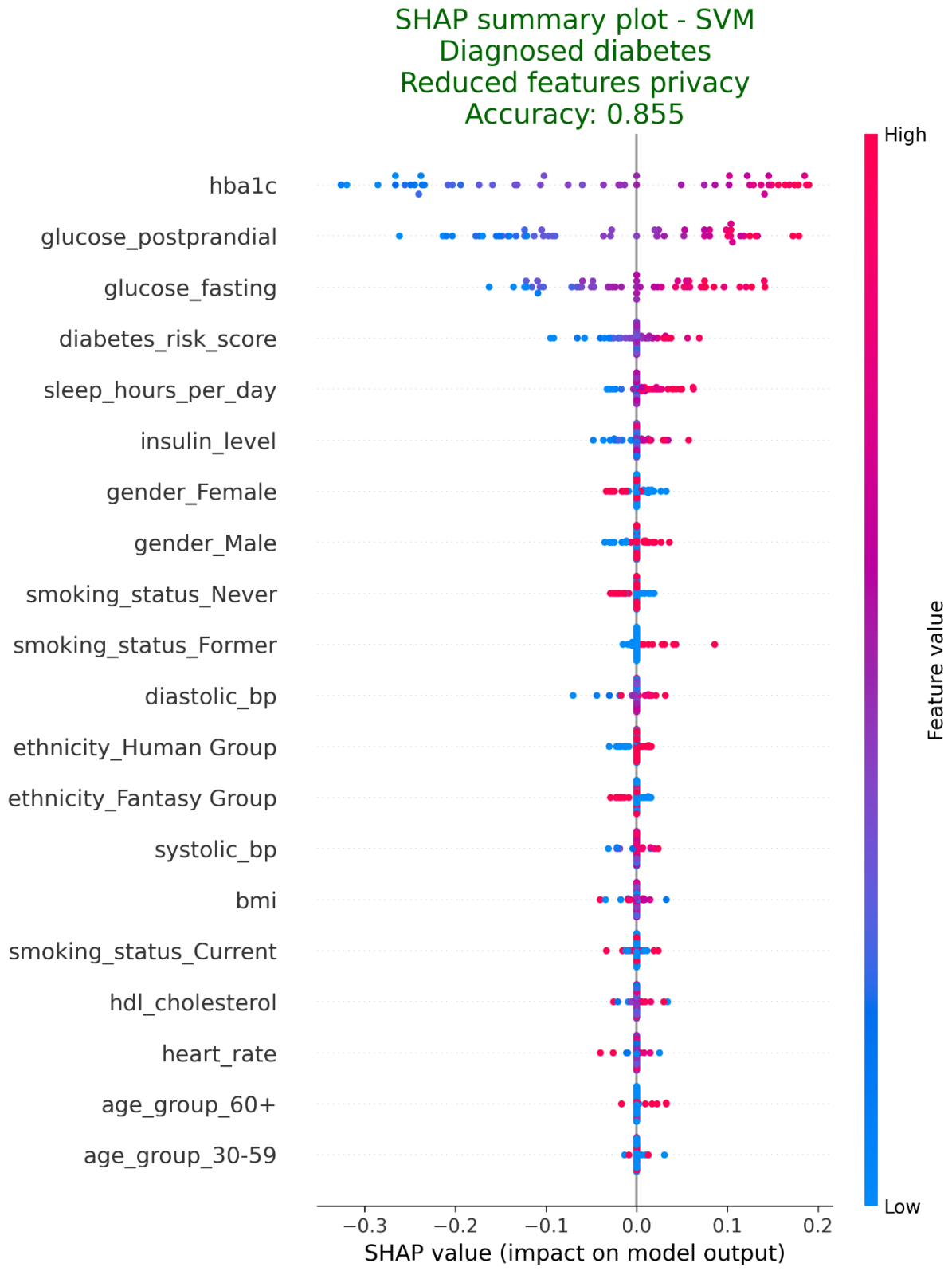


Figure 20. Support vector machines SHAP summary plot with privacy-preserving techniques applied.

For redundancy reasons, the SHAP bar plot for privacy-preserving techniques was excluded, as it presents the same data as Figure 20 in a different visual format.

5.3.3 SHAP on K-Nearest Neighbors (KNN)

Figure 21 presents the SHAP summary plot explanations for the SVMs.

Feature names are shown on the left side of the graph and include sensitive information such as ethnicity. The SHAP accuracy score for (A) is 0.785 and for (B) is 0.720.

The top five features for (A) are HbA1c, postprandial glucose, diabetes stage, fasting glucose, and gender (female).

The top five features for (B) are HbA1c, postprandial glucose, fasting glucose, gender (female), and gender (male).

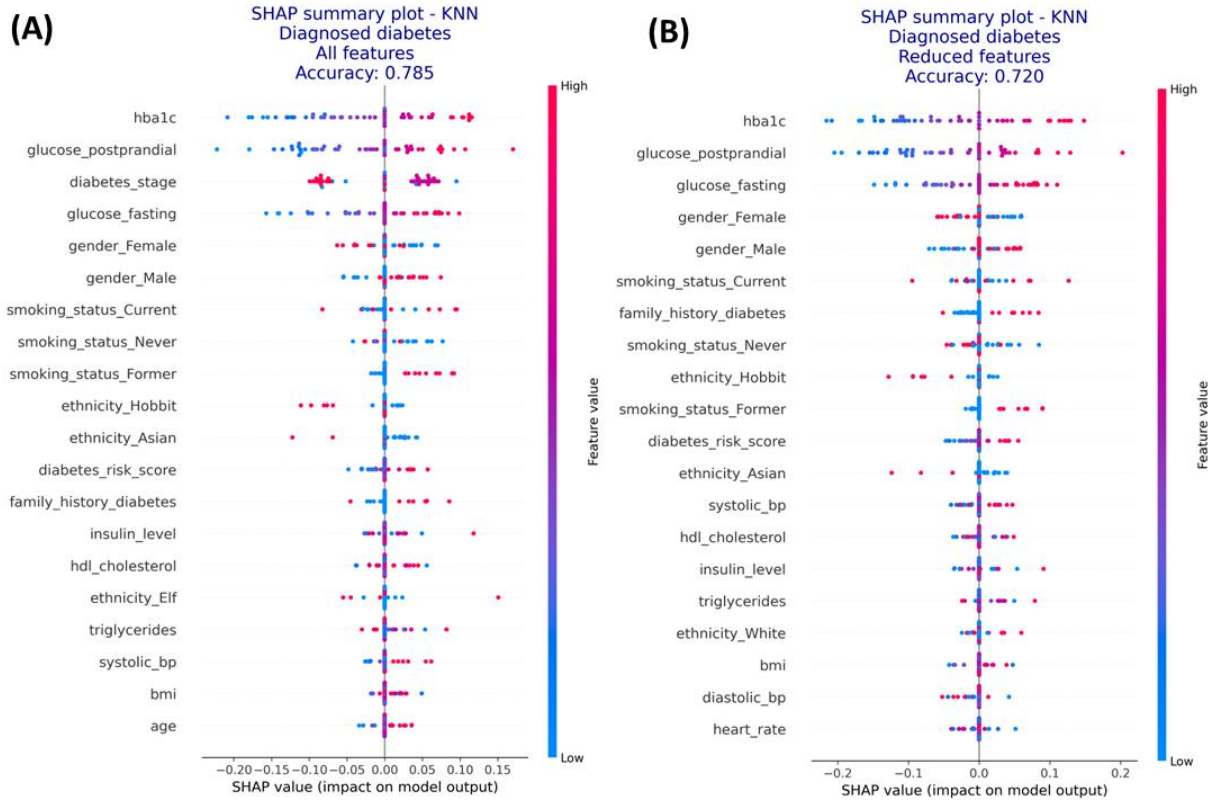


Figure 21. K-nearest neighbors SHAP summary plot feature comparison: (A) full feature set and (B) reduced feature set.

Figure 22 presents the SHAP bar plot explanations for the KNN. The results are the same as in the SHAP summary plot and only the data presentation is changed.

Feature names are shown on the left side of the graph and include sensitive information such as ethnicity. The SHAP score for (A) is 0.785 and for (B) is 0.720.

The top five features for (A) are HbA1c, postprandial glucose, diabetes stage, fasting glucose, and gender (female).

The top five features for (B) are HbA1c, postprandial glucose, fasting glucose, gender (female), and gender (male).

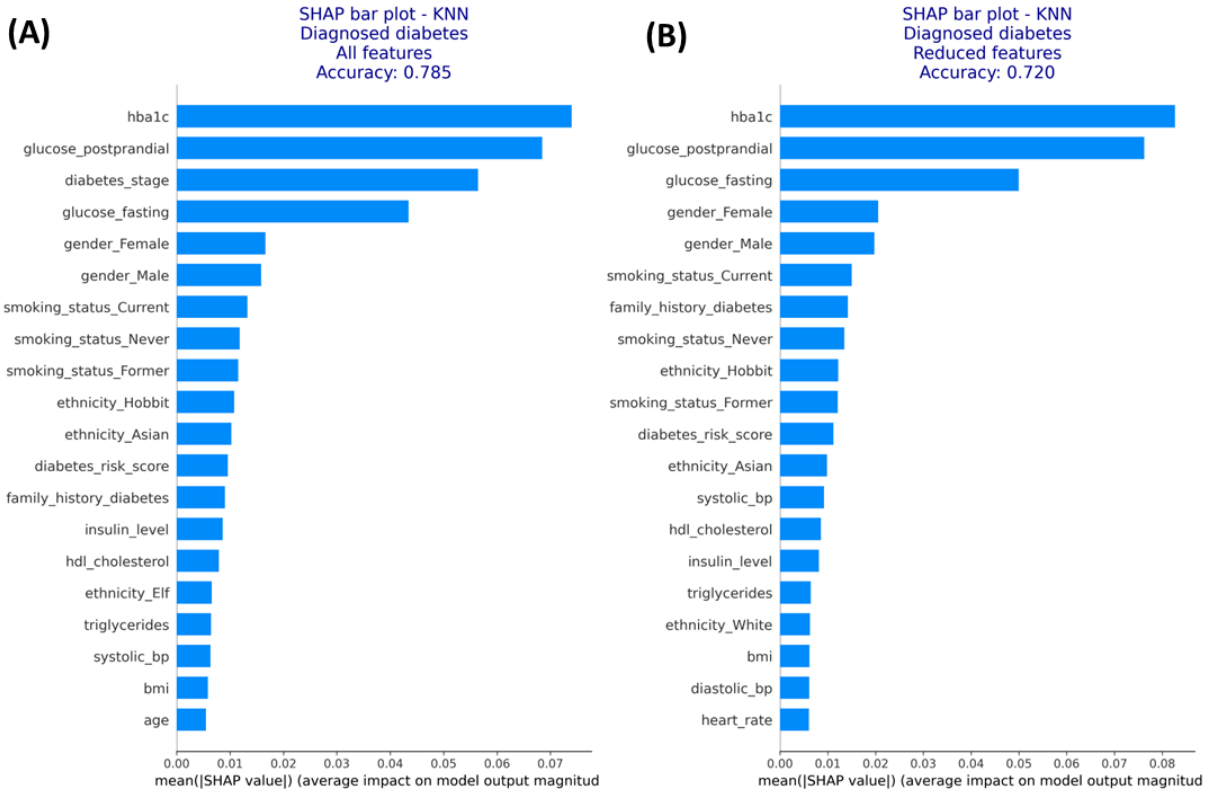


Figure 22. K-nearest neighbors SHAP bar plot feature comparison: (A) full feature set and (B) reduced feature set.

Figure 23 presents the SHAP summary plot explanations with privacy-preserving techniques applied to the KNN.

The SHAP accuracy score is 0.735.

The top five features are HbA1c, postprandial glucose, fasting glucose, former smoker, and gender (female).

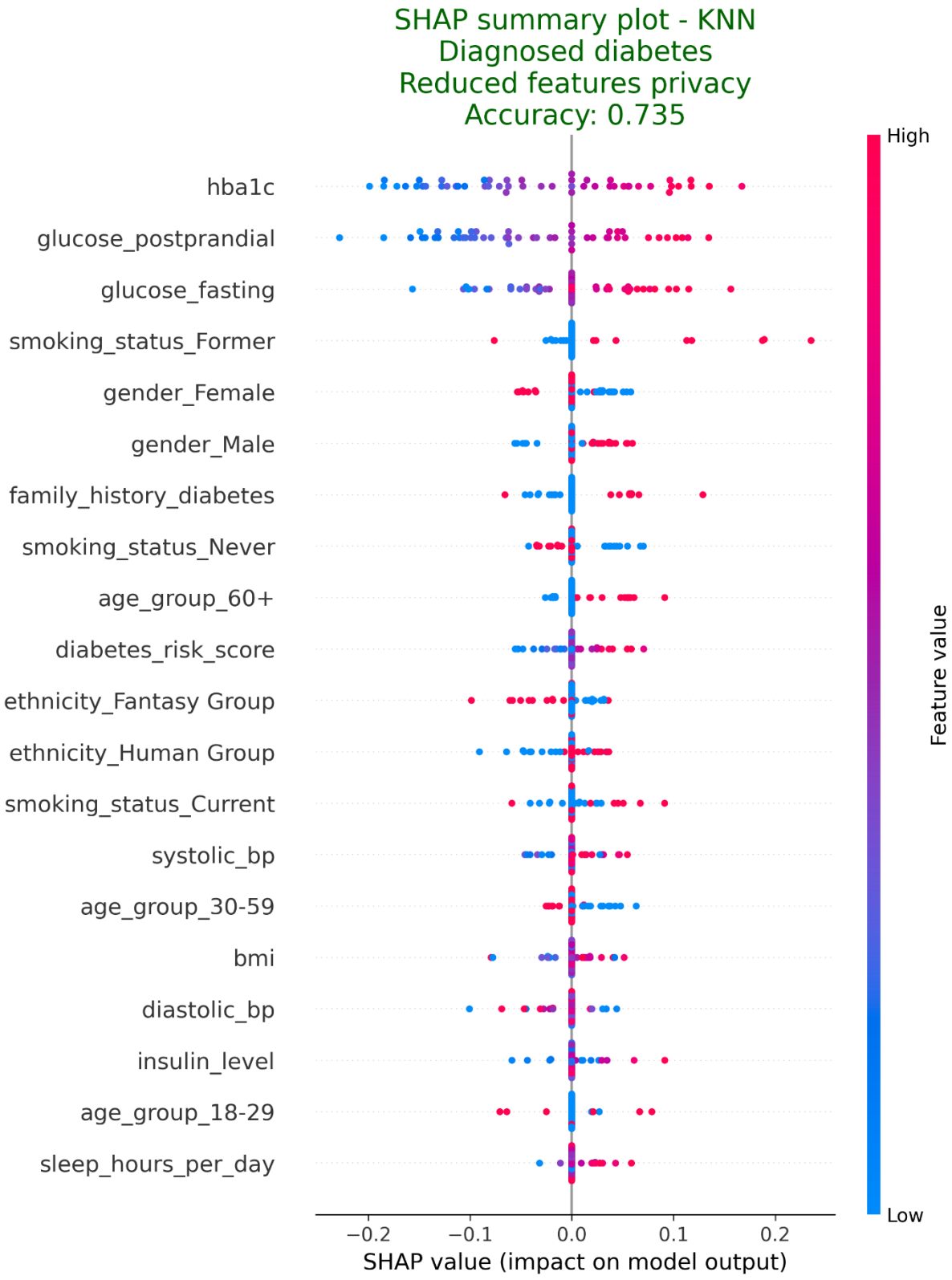


Figure 23. K-nearest neighbor SHAP summary plot with privacy-preserving techniques applied.

For redundancy reasons, the SHAP bar plot for privacy-preserving techniques was excluded, as it presents the same data as Figure 23 in a different visual format.

5.4 ELI5 Results

This section presents the results of DTs, SVMs, and KNN models using ELI5 as the explainability method.

5.4.1 ELI5 on Decision Trees (DTs)

Figure 24 presents the ELI5 explanations for the DTs. Feature names are listed in the graph and include sensitive information such as ethnicity.

The ELI5 score for (A) is 1.000 and for (B) is 0.808.

Eli5 identifies only two features for (A), diabetes stage and HbA1c.

The top five features for (B) are HbA1c, fasting glucose, systolic blood pressure, diabetes risk score, and HDL cholesterol.

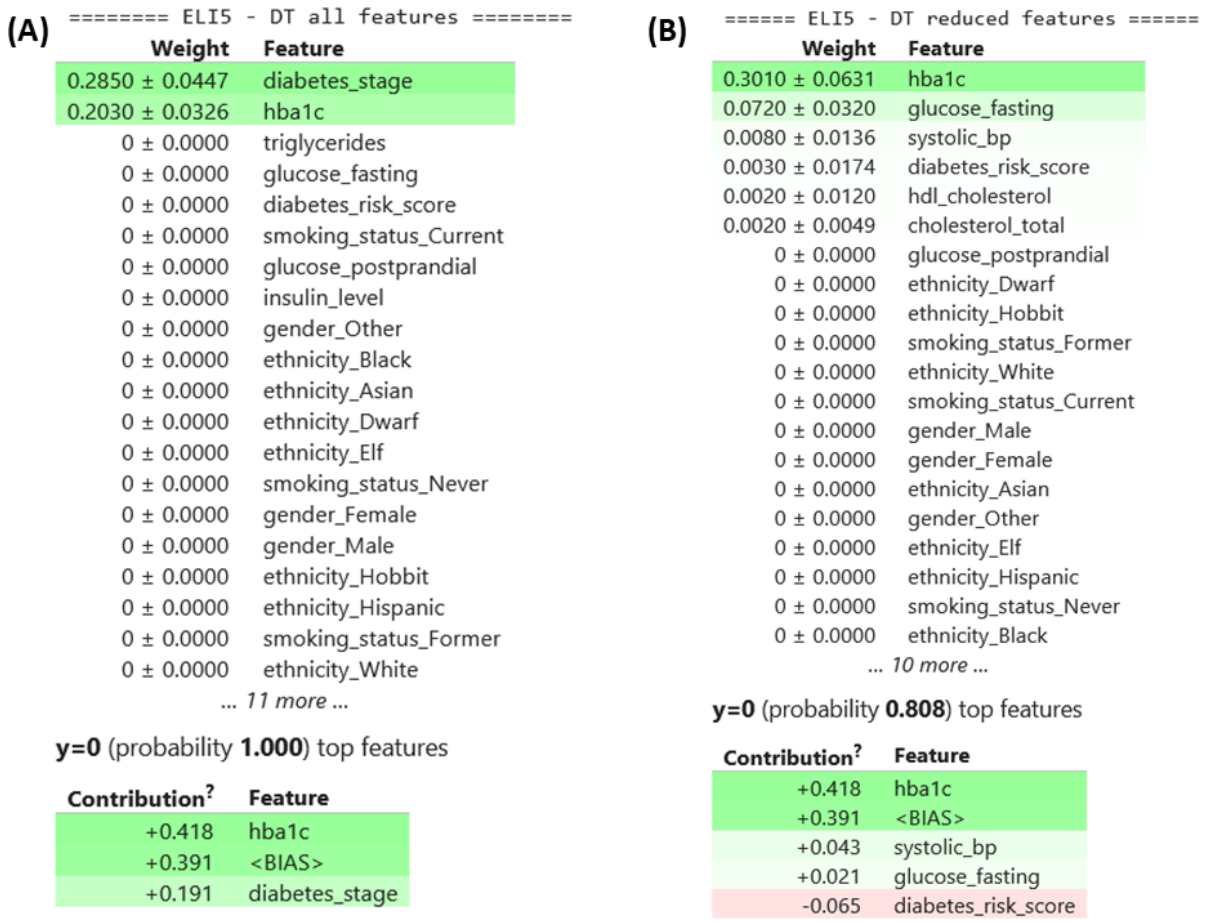


Figure 24. Decision tree ELI5 feature comparison: (A) full feature set and (B) reduced feature set.

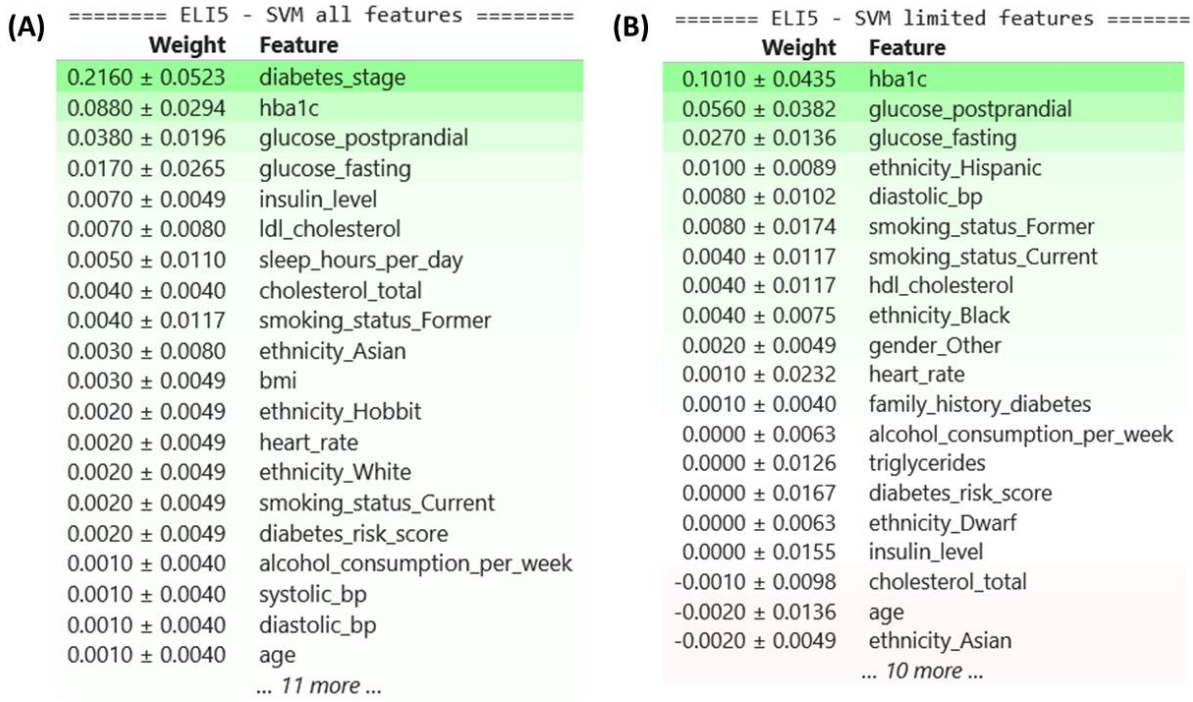
5.4.2 ELI5 on Support Vector Machines (SVMs)

Figure 25 presents the ELI5 explanations for the SVMs. Feature names are listed in the graph and include sensitive information such as ethnicity.

Eli5 scores for (A) and (B) were excluded due to compatibility limitations.

The top five features for (A) are diabetes stage, HbA1c, postprandial glucose, fasting glucose, and insulin level.

The top five features for (B) are HbA1c, postprandial glucose, fasting glucose, Hispanic ethnicity, and diastolic blood pressure.



(A)

======== ELI5 - SVM all features ========

Weight	Feature
0.2160 ± 0.0523	diabetes_stage
0.0880 ± 0.0294	hba1c
0.0380 ± 0.0196	glucose_postprandial
0.0170 ± 0.0265	glucose_fasting
0.0070 ± 0.0049	insulin_level
0.0070 ± 0.0080	ldl_cholesterol
0.0050 ± 0.0110	sleep_hours_per_day
0.0040 ± 0.0040	cholesterol_total
0.0040 ± 0.0117	smoking_status_Former
0.0030 ± 0.0080	ethnicity_Asian
0.0030 ± 0.0049	bmi
0.0020 ± 0.0049	ethnicity_Hobbit
0.0020 ± 0.0049	heart_rate
0.0020 ± 0.0049	ethnicity_White
0.0020 ± 0.0049	smoking_status_Current
0.0020 ± 0.0049	diabetes_risk_score
0.0010 ± 0.0040	alcohol_consumption_per_week
0.0010 ± 0.0040	systolic_bp
0.0010 ± 0.0040	diastolic_bp
0.0010 ± 0.0040	age
	... 11 more ...

(B)

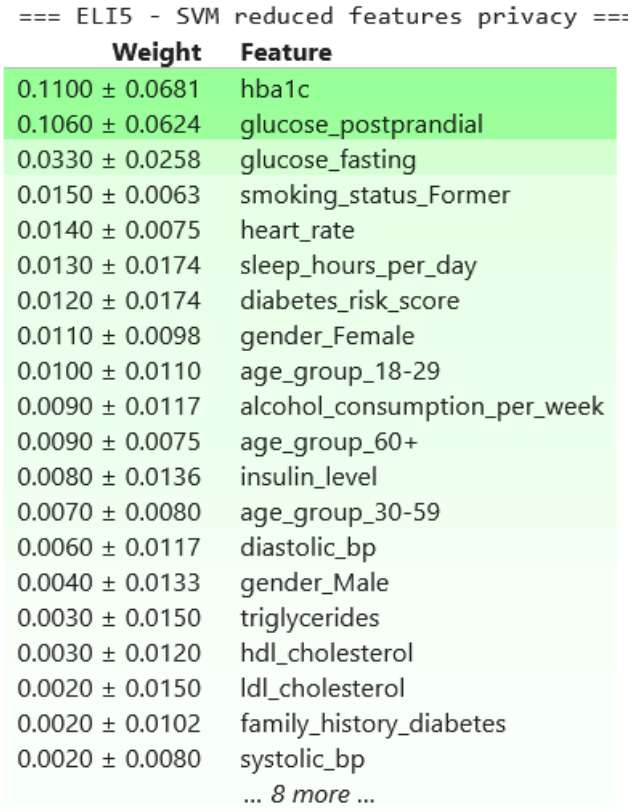
======= ELI5 - SVM limited features =======

Weight	Feature
0.1010 ± 0.0435	hba1c
0.0560 ± 0.0382	glucose_postprandial
0.0270 ± 0.0136	glucose_fasting
0.0100 ± 0.0089	ethnicity_Hispanic
0.0080 ± 0.0102	diastolic_bp
0.0080 ± 0.0174	smoking_status_Former
0.0040 ± 0.0117	smoking_status_Current
0.0040 ± 0.0117	hdl_cholesterol
0.0040 ± 0.0075	ethnicity_Black
0.0020 ± 0.0049	gender_Other
0.0010 ± 0.0232	heart_rate
0.0010 ± 0.0040	family_history_diabetes
0.0000 ± 0.0063	alcohol_consumption_per_week
0.0000 ± 0.0126	triglycerides
0.0000 ± 0.0167	diabetes_risk_score
0.0000 ± 0.0063	ethnicity_Dwarf
0.0000 ± 0.0155	insulin_level
-0.0010 ± 0.0098	cholesterol_total
-0.0020 ± 0.0136	age
-0.0020 ± 0.0049	ethnicity_Asian
	... 10 more ...

Figure 25. Support vector machines ELI5 feature comparison: (A) full feature set and (B) reduced feature set.

Figure 26 presents the ELI5 explanations with privacy-preserving techniques applied for the SVMs.

The top five features are HbA1c, postprandial glucose, fasting glucose, former smoker, and heart rate.



=== ELI5 - SVM reduced features privacy ===

Weight	Feature
0.1100 ± 0.0681	hba1c
0.1060 ± 0.0624	glucose_postprandial
0.0330 ± 0.0258	glucose_fasting
0.0150 ± 0.0063	smoking_status_Former
0.0140 ± 0.0075	heart_rate
0.0130 ± 0.0174	sleep_hours_per_day
0.0120 ± 0.0174	diabetes_risk_score
0.0110 ± 0.0098	gender_Female
0.0100 ± 0.0110	age_group_18-29
0.0090 ± 0.0117	alcohol_consumption_per_week
0.0090 ± 0.0075	age_group_60+
0.0080 ± 0.0136	insulin_level
0.0070 ± 0.0080	age_group_30-59
0.0060 ± 0.0117	diastolic_bp
0.0040 ± 0.0133	gender_Male
0.0030 ± 0.0150	triglycerides
0.0030 ± 0.0120	hdl_cholesterol
0.0020 ± 0.0150	ldl_cholesterol
0.0020 ± 0.0102	family_history_diabetes
0.0020 ± 0.0080	systolic_bp
	... 8 more ...

Figure 26. Support vector machines ELI5 summary plot with privacy-preserving techniques applied

5.4.3 ELI5 on K-Nearest Neighbors (KNN)

Figure 27 presents the ELI5 explanations for the KNN. Feature names are listed in the graph and include sensitive information such as ethnicity.

Eli5 scores for (A) and (B) were excluded due to compatibility limitations.

The top five features for (A) are diabetes stage, postprandial glucose, HbA1c, fasting glucose, and Dwarf ethnicity.

The top five features for (B) are postprandial glucose, HbA1c, fasting glucose, BMI, and Dwarf ethnicity.

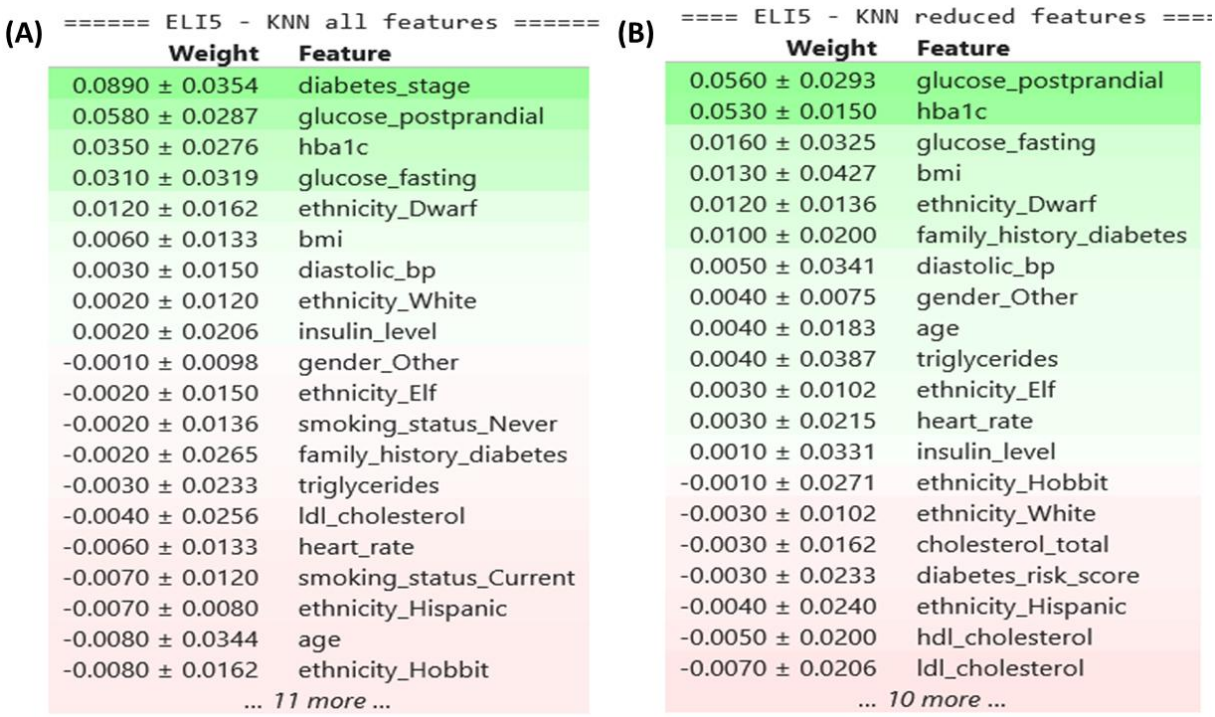


Figure 27. K-nearest neighbors ELI5 feature comparison: (A) full feature set and (B) reduced feature set.

Figure 28 presents the ELI5 explanations with privacy-preserving techniques applied for the KNN.

The top five features are HbA1c, postprandial glucose, fasting glucose, age group (60+), and gender (other).

==== ELI5 - KNN reduced features privacy

Weight	Feature
0.0740 ± 0.0264	hba1c
0.0640 ± 0.0515	glucose_postprandial
0.0200 ± 0.0200	glucose_fasting
0.0170 ± 0.0224	age_group_60+
0.0090 ± 0.0075	gender_Other
0.0020 ± 0.0344	smoking_status_Never
0.0000 ± 0.0268	ethnicity_Human Group
0.0000 ± 0.0268	cholesterol_total
-0.0030 ± 0.0196	family_history_diabetes
-0.0030 ± 0.0102	ldl_cholesterol
-0.0040 ± 0.0366	gender_Male
-0.0060 ± 0.0075	smoking_status_Current
-0.0060 ± 0.0183	insulin_level
-0.0080 ± 0.0080	bmi
-0.0090 ± 0.0117	age_group_18-29
-0.0120 ± 0.0356	gender_Female
-0.0130 ± 0.0215	ethnicity_Fantasy Group
-0.0130 ± 0.0344	age_group_30-59
-0.0140 ± 0.0440	diastolic_bp
-0.0170 ± 0.0258	triglycerides
… *7 more* …	

Figure 28. K-nearest neighbor ELI5 summary plot with privacy-preserving techniques applied.

6 Discussion

6.1 Summary of Findings

This thesis examined if privacy information would have been shown in an extensive way when applying XAI methods. Three special patients were introduced; patient A (250 years old Elf), patient B (124 years old Elf), and patient C (with gestational diabetes). None of the patients were directly identifiable from the model explanations with the configurations used.

The dataset used in this study was imbalanced, as most patients had diabetes. Because the primary focus of this work was on privacy aspects rather than evaluating diabetes prediction, the features affecting diabetes were not evaluated, and the results are not intended to be accurate for real world clinical use.

Ethnicity was a common feature in the models. For example, Dwarfs had a relatively small representation, 50 out of 1,000 patients (5%), but the XAI methods distinguished them few times in the explanations. The results demonstrate that unusual or distinguishable patient groups can become influential in model explanations even when not largely represented.

With different configurations, for example changing the target feature to “diabetes_stage”, it may be possible to reveal individuals and apply different kind of attack types, such as membership inference attacks. The entire dataset contained only 1 gestational patient; patient C. If the target feature would have been “diabetes_stage” instead of “diagnosed_diabetes”, patient C may have been distinguished.

When applying privacy-preserving methods (random noise and k-anonymization style generalization), the explainability results differ from those obtained without such methods. While noise and generalization are effective in reducing the risk of identification, they also alter the results and may introduce bias. The generalization used in this study was rough, for example age groups could have been segmented more carefully for more realistic results. In k-anonymity based approaches, a minimum group size is five records to maintain privacy. As shown in Table 6 (p. 37), after applying generalization, the age group balanced distribution became more balanced.

Balancing privacy-preserving techniques and result quality should be carefully considered, as too strict privacy protection will reduce the trustworthiness of the results, whereas too lenient protection may breach regulations and laws. Standardized frameworks are needed to mitigate

risks like model poisoning while ensuring that datasets remain both privacy-preserving and robust while maintaining explainability through XAI.

6.2 Answering Research Questions

- RQ1: What are the key cybersecurity and privacy risks associated with the use of AI in healthcare systems?

Key cybersecurity and privacy risks associated with the use of AI in healthcare systems include data breaches, unauthorized access to patient information, misuse of patient information, AI data poisoning, and privacy leakage through XAI.

In the context of the study, data poisoning can pose a significant threat, as demonstrated with the Dwarf ethnicity. The records were assigned randomly during the data preprocessing, and Dwarfs represented 5% of the total patients. Still, they were distinguished quite often by XAI methods. Modifying records showed how relatively small changes in dataset configuration can influence model behavior and the resulting explanations. This is similar to how data poisoning attacks work in practice.

- RQ2: How do XAI methods (such as LIME, SHAP, and ELI5) influence the exposure of sensitive information in healthcare datasets?

XAI methods (such as LIME, SHAP, and ELI5) can increase the exposure of sensitive information in healthcare datasets by highlighting influential features of individual or unusual patients. Smaller datasets are more likely to produce simpler results and may increase the risk of overfitting and unstable explanations. When adding XAI methods to white-box models, such as DTs, the combination of decision paths and feature importance information may reveal how specific predictions are formed and increase the risk of indirectly identifying individuals.

In the dataset, there was only one patient with gestational diabetes. Such a rare or unique feature may increase the risk of privacy exposure, as they can make individuals more easily identifiable through explanation methods. Adding random noise can help to protect small, unique groups, at the cost of reduced result credibility.

- RQ3: How do dataset characteristics (e.g., unique values, feature distribution) influence model explainability and potential privacy risks?

Unique values and feature distribution can affect how clearly ML models can be interpreted. When features contain many unique values, decision boundaries tend to become more fragmented, which may reduce the overall clarity of explanations.

Imbalanced feature distribution may cause certain patterns or groups to be more strongly represented in model explanations. This can increase the risks of privacy exposure and indirect identifiability through explanation outputs. In this study, the age feature did not have a strong importance on the explanations, and the Elf patient with an age of 250 years was not highlighted in model explanations.

6.3 Privacy and Security Implications

The results of this study suggest that privacy and security risks in AI-based healthcare systems are strongly linked. Data poisoning, unauthorized access, and misuse of sensitive information represent direct cybersecurity threats, while XAI methods introduce additional privacy risks by revealing information about individuals. Privacy and security in AI healthcare systems cannot be treated separately, and they must be considered jointly throughout data preprocessing, model development, and explainability analysis.

Additionally, the increasing use of LLMs introduces further security, privacy, and explainability considerations. The handling of sensitive patient data and risk of unintended data exposure during model training or deployment should be carefully addressed. LLMs may introduce biases and hallucinations, which can affect the trustworthiness of outputs in healthcare contexts.

6.4 Legislative Considerations

The dataset consists of synthetic patient data and therefore is not obligated to follow data protection legislation. However, if real-world data were used, several regulatory frameworks would need to be considered. As discussed earlier, the GDPR, NIS2, and AIA are strict about how personal information should be handled and failing to meet the requirements may lead to significant penalties.

Individuals have the right to request the erasure their personal data (“right to be forgotten”), and organizations handling such data should implement mechanisms to ensure compliance with such requests. However, as discussed earlier, certain patient laws may overrule the right to erasure. In the case of LLMs, right to erasure is more complex, as they are trained on

billions of parameters, and it is not always possible to determine how specific data points have influenced the learning process.

6.5 Limitations and Challenges

Limitations of the study include the relatively small size of the dataset and overrepresentation of diabetic patients. In this regard, the dataset cannot be considered representative of a real-world population. Another limitation is that the models may be biased, which can influence the classification performance.

Another limitation was computational performance. Despite the dataset being 1,000 rows from the original 100,000 rows, SHAP results were slow and computationally costly. A single SHAP prediction for SVMs took over four minutes to execute, whereas LIME produced results almost instantly.

Implementing SHAP to different ML methods proved to be challenging sometimes. Common errors included mismatch of data formats with the SHAP waterfall plot, and eventually it was excluded from the data visualization results.

Balancing with privacy-preserving techniques and dataset proved to be challenging. With more careful consideration, the data could have been distributed more evenly for more accurate results. When generalizing groups of any kind and merging them, causality is lost. Balancing research with regulations and data protection can be challenging and on a large scale some bias is bound to happen.

Minor challenges were the modification of CSV-file, since Microsoft Excel kept changing commas to semicolons due to localization settings. Excel also changed the values in columns to scientific mathematical format even when cell formatting was set to text. For this reason, Google Sheets was used for data preprocessing. However, processing data of real-life persons is not appropriate on Google's systems due to GDPR and other regulative constraints.

With more time, it would have been possible to investigate how unusual individuals may have influenced privacy exposure.

7 Conclusion

This master's thesis examined cybersecurity, privacy, and regulatory considerations related to AI and XAI in healthcare systems. The study demonstrated how relatively small and random modifications to datasets can influence model outputs as well as potential security and privacy risks. Privacy-preserving frameworks, such as X-PRISM, could become increasingly important when handling sensitive patient data with AI systems.

This thesis also presented the evolution of AI throughout the decades and its significance in the Fourth Industrial Revolution (4IR), together with related legislatives, cybersecurity threats, and privacy issues. AI is an umbrella term for intelligent systems and there are many different approaches. AI will become more ubiquitous and relevant in everyday life. In the healthcare sector, it has significant potential to advance medical research and improve patient diagnosis and treatment.

7.1 Contributions

This thesis contributes to the understanding of privacy and cybersecurity risks in AI-based healthcare systems, particularly in the context of XAI. The study combines ML with XAI methods to analyze how model explanations can affect the exposure of sensitive information.

A key contribution of this study is the demonstration of how dataset characteristics can influence both model explainability and privacy risks. The study shows how relatively small modifications in dataset configuration can affect explanation stability and potentially increase indirect identifiability of individuals.

In addition, the thesis provides an overview of relevant regulations and laws, particularly within the European Union, but also considering selected global perspectives.

7.2 Future Work

Future research could expand this study by using larger and more realistic healthcare datasets with more realistic privacy anomalies. The dataset used in this study consisted of synthetic patient records with intentionally modified values. This study applied three different ML models and three different XAI methods. With additional testing, more detailed insights into privacy exposure and explainability could be obtained. A closer inspection of deep learning methods and global regulations could also be conducted in future research.

Future research could also study the use of LLMs in healthcare systems, as these were not examined in this thesis. As LLMs become more common, further research is needed to examine their explainability, privacy implications, cybersecurity risks, and regulatory challenges when processing sensitive healthcare data.

Lastly, future studies could explore ethical and philosophical considerations related to AI. Environmental impacts are also significant, as AI systems consume vast amounts of electricity.

References

[1] "What is industry 4.0 and the Fourth Industrial Revolution? | McKinsey." Accessed: Feb. 26, 2026. [Online]. Available: https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-are-industry-4-0-the-fourth-industrial-revolution-and-4ir#/

[2] iED Team, "A Brief History of The 4 Industrial Revolutions that Shaped the World," Institute of Entrepreneurship Development. Accessed: Feb. 26, 2026. [Online]. Available: https://ied.eu/project-updates/the-4-industrial-revolutions/

[3] "What Is Artificial Intelligence (AI)? | IBM." Accessed: Mar. 09, 2026. [Online]. Available: https://www.ibm.com/think/topics/artificial-intelligence

[4] N. Networks, "What is the Fourth Industrial Revolution (4IR)?," Neos Networks. Accessed: Feb. 26, 2026. [Online]. Available: https://neosnetworks.com/resources/blog/what-is-fourth-industrial-revolution-4ir/

[5] "Why Is AI Bad? Artifical Intelligence's Dark Side Explained." Accessed: Apr. 15, 2026. [Online]. Available: https://naps.edu.au/blog/artificial-intelligence-ai-the-bad

[6] R. D. Caballar, "10 AI dangers and risks and how to manage them | IBM." Accessed: Apr. 15, 2026. [Online]. Available: https://www.ibm.com/think/insights/10-ai-dangers-and-risks-and-how-to-manage-them

[7] T. P. Quinn, S. Jacobs, M. Senadeera, V. Le, and S. Coghlan, "The three ghosts of medical AI: Can the black-box present deliver?," *Artif. Intell. Med.*, vol. 124, p. 102158, Feb. 2022, doi: 10.1016/j.artmed.2021.102158.

[8] "Ethics guidelines for trustworthy AI | Shaping Europe's digital future." Accessed: May 09, 2026. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[9] "Implementation Timeline | EU Artificial Intelligence Act." Accessed: May 09, 2026. [Online]. Available: https://artificialintelligenceact.eu/implementation-timeline/

[10] "Turing machine | Definition & Facts | Britannica." Accessed: May 16, 2026. [Online]. Available: https://www.britannica.com/technology/Turing-machine

[11] J. McCarthy, "WHAT IS ARTIFICIAL INTELLIGENCE?," Nov. 2007, Accessed: Mar. 12, 2026. [Online]. Available: https://www-formal.stanford.edu/jmc/whatisai.pdf

[12] N. Berente, B. Gu, J. Recker, and R. Santhanam, "Managing Artificial Intelligence," *MIS Q.*, vol. 45, pp. 1433–1450, Sep. 2021, doi: 10.25300/MISQ/2021/16274.

[13] D. Bergmann, “What is Machine Learning? | IBM.” Accessed: Apr. 19, 2026. [Online]. Available: https://www.ibm.com/think/topics/machine-learning

[14] T. M. Stryker Cole, “What Is Artificial Superintelligence? | IBM.” Accessed: Apr. 12, 2026. [Online]. Available: https://www.ibm.com/think/topics/artificial-superintelligence

[15] “Types of Artificial Intelligence | IBM.” Accessed: Mar. 24, 2026. [Online]. Available: https://www.ibm.com/think/topics/artificial-intelligence-types

[16] I. El Naqa and M. J. Murphy, “What Is Machine Learning?,” in *Machine Learning in Radiation Oncology: Theory and Applications*, I. El Naqa, R. Li, and M. J. Murphy, Eds., Cham: Springer International Publishing, 2015, pp. 3–11. doi: 10.1007/978-3-319-18305-3_1.

[17] “What is the difference between labeled and unlabeled data?,” GeeksforGeeks. Accessed: Apr. 22, 2026. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/what-is-the-difference-between-labeled-and-unlabeled-data/

[18] “What Is Data Labeling? | IBM.” Accessed: Apr. 22, 2026. [Online]. Available: https://www.ibm.com/think/topics/data-labeling

[19] D. Bergmann, “What Are Machine Learning Algorithms? | IBM.” Accessed: Apr. 19, 2026. [Online]. Available: https://www.ibm.com/think/topics/machine-learning-algorithms

[20] “Types of Machine Learning - Javatpoint | PDF,” Scribd. Accessed: Apr. 22, 2026. [Online]. Available: https://www.scribd.com/document/1019680293/Types-of-Machine-Learning-Javatpoint

[21] “What Is Unsupervised Learning? | IBM.” Accessed: Mar. 24, 2026. [Online]. Available: https://www.ibm.com/think/topics/unsupervised-learning

[22] A. Thombre, “Explainable AI (XAI): Using decision trees to explain neural network model,” Sep. 2024.

[23] “Decision Tree,” GeeksforGeeks. Accessed: May 05, 2026. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/decision-tree/

[24] “1.10. Decision Trees,” scikit-learn. Accessed: May 06, 2026. [Online]. Available: https://scikit-learn/stable/modules/tree.html

[25] E. Kavlakoglu, “What Is Support Vector Machine? | IBM.” Accessed: May 08, 2026. [Online]. Available: https://www.ibm.com/think/topics/support-vector-machine

[26] “Support Vector Machine (SVM) Algorithm,” GeeksforGeeks. Accessed: May 08, 2026. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/support-vector-machine-algorithm/

[27] “K-Nearest Neighbor(KNN) Algorithm,” GeeksforGeeks. Accessed: May 08, 2026. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/k-nearest-neighbours/

[28] E. Kavlakoglu, “What is the k-nearest neighbors algorithm? | IBM.” Accessed: May 08, 2026. [Online]. Available: https://www.ibm.com/think/topics/knn

[29] “What Is Deep Learning? | IBM.” Accessed: Mar. 24, 2026. [Online]. Available: https://www.ibm.com/think/topics/deep-learning

[30] C. S. Holdsworth Jim, “What Is NLP (Natural Language Processing)? | IBM.” Accessed: Apr. 19, 2026. [Online]. Available: https://www.ibm.com/think/topics/natural-language-processing

[31] C. S. Bergmann Dave, “What is a Transformer Model? | IBM.” Accessed: Apr. 19, 2026. [Online]. Available: https://www.ibm.com/think/topics/transformer-model

[32] C. Stryker, “What Are Large Language Models (LLMs)? | IBM.” Accessed: Apr. 10, 2026. [Online]. Available: https://www.ibm.com/think/topics/large-language-models

[33] Owaspg. Editor, “LLM04:2025 Data and Model Poisoning,” OWASP Gen AI Security Project. Accessed: Apr. 10, 2026. [Online]. Available: https://genai.owasp.org/llmrisk/llm04-model-denial-of-service/

[34] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman, and S. Rab, “Significance of machine learning in healthcare: Features, pillars and applications,” *Int. J. Intell. Netw.*, vol. 3, pp. 58–73, Jan. 2022, doi: 10.1016/j.ijin.2022.05.002.

[35] Q. An, S. Rahman, J. Zhou, and J. J. Kang, “A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges,” *Sensors*, vol. 23, no. 9, p. 4178, Jan. 2023, doi: 10.3390/s23094178.

[36] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, “Large language models in medicine,” *Nat. Med.*, vol. 29, no. 8, pp. 1930–1940, Aug. 2023, doi: 10.1038/s41591-023-02448-8.

[37] J. Brickman, M. Gupta, and J. R. Oltmanns, “Large Language Models for Psychological Assessment: A Comprehensive Overview,” *Adv. Methods Pract. Psychol. Sci.*, vol. 8, no. 3, p. 25152459251343582, Jul. 2025, doi: 10.1177/25152459251343582.

[38] R. Dwivedi *et al.*, “Explainable AI (XAI): Core Ideas, Techniques, and Solutions,” *ACM Comput Surv*, vol. 55, no. 9, p. 194:1-194:33, Jan. 2023, doi: 10.1145/3561048.

[39] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 3rd ed., vol. 2025. Accessed: Apr. 21, 2026. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[40] F. Ezzeddine, “Privacy Implications of Explainable AI in Data-Driven Systems,” Jun. 22, 2024, *arXiv*: arXiv:2406.15789. doi: 10.48550/arXiv.2406.15789.

[41] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD ’16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[42] V. Vishwarupe, P. M. Joshi, N. Mathias, S. Maheshwari, S. Mhaisalkar, and V. Pawar, “Explainable AI and Interpretable Machine Learning: A Case Study in Perspective,” *Procedia Comput. Sci.*, vol. 204, pp. 869–876, 2022, doi: 10.1016/j.procs.2022.08.105.

[43] I. D. Mienye *et al.*, “A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges,” *Inform. Med. Unlocked*, vol. 51, p. 101587, Jan. 2024, doi: 10.1016/j.imu.2024.101587.

[44] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” Nov. 25, 2017, *arXiv*: arXiv:1705.07874. doi: 10.48550/arXiv.1705.07874.

[45] I. D. Mienye and N. Jere, “Optimized Ensemble Learning Approach with Explainable AI for Improved Heart Disease Prediction,” *Information*, vol. 15, no. 7, p. 394, Jul. 2024, doi: 10.3390/info15070394.

[46] “Welcome to ELI5’s documentation! — ELI5 0.16.0 documentation.” Accessed: May 08, 2026. [Online]. Available: https://eli5.readthedocs.io/en/stable/

[47] L. Kost, S. K. Lier, and M. H. Breitner, “An explainable artificial intelligence feature selection framework for transparent, trustworthy, and cost-efficient energy forecasting,” *Energy AI*, vol. 22, p. 100648, Dec. 2025, doi: 10.1016/j.egyai.2025.100648.

[48] “Brussels launched an age checking app. Hackers say it takes 2 minutes to break it.,” POLITICO. Accessed: Apr. 22, 2026. [Online]. Available: https://www.politico.eu/article/eu-brussels-launched-age-checking-app-hackers-say-took-them-2-minutes-break-it/

[49] S. Li, K. Surineni, and N. Prabhakaran, “Cyber-Attacks on Hospital Systems: A Narrative Review,” *Am. J. Geriatr. Psychiatry Open Sci. Educ. Pract.*, vol. 7, pp. 30–39, Sep. 2025, doi: 10.1016/j.osep.2025.03.002.

[50] “Types of legislation | European Union.” Accessed: May 02, 2026. [Online]. Available: https://european-union.europa.eu/institutions-law-budget/law/types-legislation_en

[51] “General data protection regulation (GDPR) | EUR-Lex.” Accessed: Apr. 12, 2026. [Online]. Available: https://eur-lex.europa.eu/EN/legal-content/summary/general-data-protection-regulation-gdpr.html

[52] jhofmann@aurorait.com, “How GDPR, CCPA, HIPAA, and Other Data Privacy Standards Safeguard Our Digital Lives,” Plurilock. Accessed: Apr. 12, 2026. [Online]. Available: https://plurilock.com/blog/how-gdpr-ccpa-hipaa-and-other-data-privacy-standards-safeguard-our-digital-lives/

[53] “GDPR Archives,” GDPR.eu. Accessed: Apr. 12, 2026. [Online]. Available: https://gdpr.eu/tag/gdpr/

[54] DPM, “20 biggest GDPR fines so far [2025],” Data Privacy Manager. Accessed: Apr. 12, 2026. [Online]. Available: https://dataprivacymanager.net/5-biggest-gdpr-fines-so-far-2020/

[55] “California Imposes Largest CCPA Fine to Date on Disney | News & Events | Clark Hill PLC.” Accessed: Apr. 12, 2026. [Online]. Available: https://www.clarkhill.com/news-events/news/california-imposes-largest-ccpa-fine-to-date-on-disney/

[56] “Global Data Privacy Laws 2026: 100+ Regulations,” CDP.com. Accessed: Apr. 12, 2026. [Online]. Available: https://cdp.com/basics/international-u-s-data-privacy-laws-and-regulations-you-need-to-know/

[57] B. Clifton, “What Is PII versus Personal Data?,” Brian Clifton’s Thoughts. Accessed: May 08, 2026. [Online]. Available: https://brianclifton.com/blog/2018/05/21/gdpr-request-consent-before-tracking/

[58] M. Thuret-Benoist, “The difference between PII and Personal Data - blog,” TechGDPR. Accessed: May 08, 2026. [Online]. Available: https://techgdpr.com/blog/difference-between-pii-and-personal-data/

[59] “Data protection explained - European Commission.” Accessed: May 08, 2026. [Online]. Available: https://commission.europa.eu/law/law-topic/data-protection/data-protection-explained_en

[60] “Health care,” Tietosuojavaltuutetun toimisto. Accessed: Apr. 22, 2026. [Online]. Available: https://tietosuoja.fi/en/faq-health-care

[61] “What is NIS2?,” The NIS2 Directive. Accessed: Apr. 23, 2026. [Online]. Available: https://nis2directive.eu/what-is-nis2/

[62] “List of NIS 2 sectors,” SPAC Alliance. Accessed: Apr. 23, 2026. [Online]. Available: https://spac-alliance.org/library/online-nis-2-guide/nis-2-sectors/

[63] “EU cybersecurity policies | Shaping Europe’s digital future.” Accessed: May 01, 2026. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/cybersecurity-policies

[64] “NIS2 Directive Transposition Tracker,” ECSO. Accessed: Apr. 23, 2026. [Online]. Available: https://ecs-org.eu/activities/nis2-directive-transposition-tracker/

[65] “Important information on the European Union Cybersecurity Directive (NIS2),” Traficom. Accessed: Apr. 23, 2026. [Online]. Available: https://www.kyberturvallisuuskeskus.fi/en/our-activities/regulation-and-supervision/nis2-european-union-cybersecurity-directive/important-information-european-union-cybersecurity-directive-nis2#67853-0

[66] “European AI Office | Shaping Europe’s digital future.” Accessed: May 09, 2026. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/ai-office

[67] “AI Act | Shaping Europe’s digital future.” Accessed: May 02, 2026. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[68] “AI Act | Shaping Europe’s digital future.” Accessed: May 09, 2026. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[69] “Who we are | ENISA.” Accessed: May 01, 2026. [Online]. Available: https://www.enisa.europa.eu/about-enisa/who-we-are

[70] I. Bala, I. Pindoo, M. Mijwil, M. Abotaleb, and W. Yundong, “Ensuring Security and Privacy in Healthcare Systems: A Review Exploring Challenges, Solutions, Future Trends, and the Practical Applications of Artificial Intelligence,” *Jordan Med. J.*, Jul. 2024, doi: https://doi.org/10.35516/jmj.v58i2.2527.

[71] SentinelOne, “Cybersecurity in Healthcare: Risks & Best Practices,” SentinelOne. Accessed: May 09, 2026. [Online]. Available: https://www.sentinelone.com/cybersecurity-101/cybersecurity/cybersecurity-in-healthcare/

[72] L. Tuure, “Snooping of Personal Data: Who Is Responsible — Organisation or Employee?,” Hannes Snellman. Accessed: May 09, 2026. [Online]. Available:

https://www.hannessnellman.com/news-and-views/blog/snooping-of-personal-data-who-is-responsible-organisation-or-employee/

[73] "Administrative fine imposed on psychotherapy centre Vastaamo for data protection violations | European Data Protection Board." Accessed: May 09, 2026. [Online]. Available: https://www.edpb.europa.eu/news/national-news/2022/administrative-fine-imposed-psychotherapy-centre-vastaamo-data-protection_en

[74] "Vastaamo victims' lawyer: Some took their own lives after patient record leak," News. Accessed: May 09, 2026. [Online]. Available: https://yle.fi/a/74-20077285

[75] "Cybersecurity of hospitals and healthcare providers | Shaping Europe's digital future." Accessed: May 01, 2026. [Online]. Available: https://digital-strategy.ec.europa.eu/en/factpages/cybersecurity-hospitals-and-healthcare-providers

[76] M. Nankya, A. Mugisa, Y. Usman, A. Upadhyay, and R. Chataut, "Security and Privacy in E-Health Systems: A Review of AI and Machine Learning Techniques," *IEEE Access*, vol. 12, pp. 148796–148816, 2024, doi: 10.1109/ACCESS.2024.3469215.

[77] M. M. Rahman, A. Siddika Arshi, M. M. Hasan, S. Farzana Mishu, H. Shahriar, and F. Wu, "Security Risk and Attacks in AI: A Survey of Security and Privacy," in *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, Jun. 2023, pp. 1834–1839. doi: 10.1109/COMPSAC57700.2023.00284.

[78] J. Holdsworth, "What Is AI Bias? | IBM." Accessed: May 03, 2026. [Online]. Available: https://www.ibm.com/think/topics/ai-bias

[79] T. K. Jonker Alexandra, "What Is Data Poisoning? | IBM." Accessed: May 09, 2026. [Online]. Available: https://www.ibm.com/think/topics/data-poisoning

[80] R. Shokri, M. Strobel, and Y. Zick, "On the Privacy Risks of Model Explanations," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, in AIES '21. New York, NY, USA: Association for Computing Machinery, Jul. 2021, pp. 231–241. doi: 10.1145/3461702.3462533.

[81] S. Allana, M. Kankanhalli, and R. Dara, "Privacy Risks and Preservation Methods in Explainable Artificial Intelligence: A Scoping Review," Dec. 02, 2025, *arXiv*: arXiv:2505.02828. doi: 10.48550/arXiv.2505.02828.

[82] A. Marques, B. Sá, R. Botelho, and P. Pinto, "Training Machine Learning Models on Encrypted Data: A Privacy-Preserving Framework using Homomorphic Encryption," Apr. 25, 2026, *arXiv*: arXiv:2604.23245. doi: 10.48550/arXiv.2604.23245.

[83] J. Trivedi, J. Isoaho, and T. Mohammad, “Enhancing Privacy Transparency in Remote Patient Monitoring with Explainable AI,” *Procedia Comput. Sci.*, vol. 265, pp. 149–156, 2025, doi: 10.1016/j.procs.2025.07.167.

[84] R. D. C. Stryker Cole, “What Is Federated Learning? | IBM.” Accessed: May 01, 2026. [Online]. Available: https://www.ibm.com/think/topics/federated-learning

[85] E.-H. Qazi, W. K. AL-Ghanem, M. H. Faheem, and H. Ullah, “Federated Learning Framework for Privacy-Preserving Explainable AI-Driven Clinical Decision-Making,” *IEEE J. Biomed. Health Inform.*, pp. 1–16, 2026, doi: 10.1109/JBHI.2026.3679499.

[86] “K-ANONYMITY AS A PRIVACY MEASURE.” Agencia Española Protección Datos Technological Surveys and Assessment Unit, May 14, 2019. Accessed: May 24, 2026. [Online]. Available: https://www.aepd.es/guides/k-anonymity-as-a-privacy-measure.pdf

[87] K. El Emam and F. K. Dankar, “Protecting Privacy Using k-Anonymity,” *J. Am. Med. Inform. Assoc. JAMIA*, vol. 15, no. 5, pp. 627–637, 2008, doi: 10.1197/jamia.M2716.

[88] R. Kolipaka and R. K. Digutla, “Diabetes Health Indicators Dataset.” Accessed: Apr. 18, 2026. [Online]. Available: https://www.kaggle.com/datasets/mohankrishnathalla/diabetes-health-indicators-dataset

[89] “Diabetes.” Accessed: May 02, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[90] “Getting Started,” scikit-learn. Accessed: Apr. 18, 2026. [Online]. Available: https://scikit-learn/stable/getting_started.html

Appendices

Appendix A: Code Used in the Research

ChatGPT (OpenAI, GPT-5 series) was used to assist in generating machine learning code examples for experimentation purposes. The prompts included requests for machine learning scripts related to the ML and XAI methods used in this thesis.

The generated code was reviewed, modified, tested, and validated by the author before use in this thesis.

```
# ============================
# DTs, SVMs, KNN with
# LIME + SHAP + ELI5
# ============================
# IMPORTS
# ============================
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import random

# KNN
# from sklearn.neighbors import KNeighborsClassifier
# from sklearn.preprocessing import StandardScaler

# SVMs
# from sklearn.svm import SVC
# from sklearn.preprocessing import StandardScaler

from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

import lime
import shap
import eli5

from lime import lime_tabular
from eli5.sklearn import PermutationImportance
from IPython.display import display

# ============================================================
# 1. LOAD DATA
# ============================================================
dataset = pd.read_csv(r"path/to/data.csv")

'''
# 1.5 PRIVACY TRANSFORMATIONS
# Used only when applying privacy techniques. Remove comments when needed
# SEED (42)
np.random.seed(42)
random.seed(42)
```

```
# NOISE INJECTION (AGE)
dataset["age"] = dataset["age"] + np.random.normal(0, 5, size=len(dataset))

# AGE GENERALIZATION
def generalize_age(x):
    if x < 30:
        return "18-29"
    elif x < 60:
        return "30-59"
    else:
        return "60+"
dataset["age_group"] = dataset["age"].apply(generalize_age)
# ETHNICITY GENERALIZATION
def generalize_ethnicity(x):
    if x in ["White", "Asian", "Black", "Hispanic"]:
        return "Human Group"
    elif x in ["Hobbit", "Elf", "Dwarf"]:
        return "Fantasy Group"
    else:
        return "generalization error"
dataset["ethnicity"] = dataset["ethnicity"].apply(generalize_ethnicity)
'''
# ============================================================
# 2. PREPROCESSING
# ============================================================
data_encoded = pd.get_dummies(dataset)

# Drop columns
# For reduced config also drop "diabetes_stage"
# For privacy config also drop "age", as it's generalized
X = data_encoded.drop(columns=["patient_id", "diagnosed_diabetes"])

# Target variable
y = data_encoded["diagnosed_diabetes"]

# ============================================================
# 3.A TRAIN / TEST SPLIT (DTs)
# ============================================================
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# 3.B SCALING for KNN and SVMs (KNN) (SVMs)
# scaler = StandardScaler()
# X_train_scaled = scaler.fit_transform(X_train)
# X_test_scaled = scaler.transform(X_test)

# ============================================================
# 4.A TRAIN DECISION TREE MODEL (DTs)
# ============================================================
clf = DecisionTreeClassifier(max_depth=None, random_state=42)
clf.fit(X_train, y_train)

# 4.B TRAIN KNN MODEL (KNN)
# knn_model = KNeighborsClassifier(n_neighbors=5)
# knn_model.fit(X_train_scaled, y_train)

# 4.C TRAIN SVMs MODEL (SVMs)
# svm_model = SVC(probability=True, random_state=42)
# svm_model.fit(X_train_scaled, y_train)
```

```
# ============================================================
# 5. EVALUATION
# ============================================================
y_pred = clf.predict(X_test)
# y_pred = knn_model.predict(X_test_scaled)
# y_pred = svm_model.predict(X_test_scaled)

# 5.5 DECISION TREE VISUALIZATION, only used for DTs
plt.figure(figsize=(6, 5)) # or values (40, 30)
plot_tree(clf, filled=True, feature_names=X.columns, class_names=["No Diabetes",
"Diabetes"],fontsize=6)

# Add titles to all where needed. Titles removed from rest of the script to reduce
redundancy
plt.title(f"DT – all features (acc: {accuracy_score(y_test, y_pred):.3f})",
fontsize=16, color="darkblue") # or darkgreen

# Save figure, add to all where needed
plt.savefig("dt_all_features_tree.png", dpi=300, bbox_inches="tight")
plt.show()

# ============================================================
# 6. LIME
# ============================================================
lime_explainer =
lime_tabular.LimeTabularExplainer(training_data=X_train.values,feature_names=X_tra
in.columns.tolist(), class_names=["No Diabetes", "Diabetes"],
mode="classification")
lime_exp = lime_explainer.explain_instance(X_test.iloc[0].values,
clf.predict_proba, num_features=10)
lime_exp.as_pyplot_figure()

# ============================================================
# 7. SHAP
# ============================================================
explainer = shap.TreeExplainer(clf)
X_sample = X_test.iloc[:100]
shap_values = explainer.shap_values(X_sample)
# Binary classification handling
if isinstance(shap_values, list):
    s_vals = shap_values[1]
else:
    s_vals = shap_values

# KNN
'''
# smaller background for speed (KNN)
X_bg = X_train_scaled[:50]
X_test_small = X_test_scaled[:50]
explainer = shap.KernelExplainer(knn_model.predict_proba, X_bg)
shap_values = explainer.shap_values(X_test_small)
if isinstance(shap_values, list):
    s_vals = shap_values[1]
else:
    s_vals = shap_values
s_vals = np.array(s_vals)
# handle 3D case
if len(s_vals.shape) == 3:
```

```
    s_vals = s_vals[:, :, 1]
s_vals = s_vals.reshape(X_test_small.shape[0], X_test_small.shape[1])
'''
# SVMs
'''
# MORE DATA, better stability
X_bg = X_train_scaled[:100]
X_test_small = X_test_scaled[:50]
explainer = shap.KernelExplainer(svm_model.predict_proba, X_bg)
shap_values = explainer.shap_values(X_test_small)

if isinstance(shap_values, list):
    s_vals = shap_values[1]
else:
    s_vals = shap_values
s_vals = np.array(s_vals)
# handle possible 3D output
if len(s_vals.shape) == 3:
    s_vals = s_vals[:, :, 1]
# enforce correct shape
s_vals = s_vals.reshape(X_test_small.shape[0], X_test_small.shape[1])
'''
# SHAP SUMMARY PLOT
shap.summary_plot(s_vals, X_sample, feature_names=X.columns, max_display=15,
show=False)
# SHAP BAR PLOT
shap.summary_plot(s_vals, X_sample, feature_names=X.columns, max_display=15,
plot_type="bar", show=False)

# ============================================================
# 8. ELI5
# ============================================================
print("\n======== ELI5 - DT all features ========")
perm = PermutationImportance(clf, random_state=42)
perm.fit(X_test, y_test)

# ELI5 FEATURE IMPORTANCE
eli5_weights = eli5.show_weights(perm,feature_names=X.columns.tolist())
display(eli5_weights)

# ELI5 SINGLE PREDICTION
eli5_prediction = eli5.show_prediction(clf, X_test.iloc[0],
feature_names=X.columns.tolist())
display(eli5_prediction)
```

Appendix B: Dataset Example

First 15 rows of the modified dataset used. The original dataset was obtained from Kaggle [88]. A total of 300 records were randomly selected, and their ethnicity values were replaced with fantasy categories (Hobbit, Elf, Dwarf). The labels are described in Table 4.

1,250,Male,Elf,Current,15,4,0,42.5,134,78,68,239,41,160,145,136,236,6.36,8.18,29.6,1,1

2,124,Female,Elf,Former,1,6.5,0,23.1,129,76,67,116,55,50,30,93,150,2,5.63,23,0,0

3,60,Male,Dwarf,Never,1,10,1,22.2,115,73,74,213,66,99,36,118,195,5.07,7.51,44.7,2,1

4,74,Female,Black,Never,0,6.6,0,26.8,120,93,68,171,50,79,140,139,253,5.28,9.03,38.2,2,1

5,46,Male,White,Never,1,7.4,0,21.2,92,67,67,210,52,125,160,137,184,12.74,7.2,23.5,2,1

6,46,Female,White,Never,2,6.2,0,26.1,95,81,57,218,61,119,179,100,133,8.77,6.03,23.5,3,0

7,75,Female,White,Never,0,7.8,0,25.1,129,77,81,238,46,161,155,101,100,10.14,5.24,36.1,3,0

8,62,Male,Elf,Current,1,9,0,23.9,128,83,76,241,49,159,120,110,189,8.96,7.04,34.2,2,1

9,42,Male,Black,Current,1,8.5,0,24.7,103,71,72,187,33,132,98,116,172,5.7,6.9,26.7,2,1

10,59,Female,Elf,Current,3,5.3,0,26.7,124,81,70,188,52,103,104,76,109,4.49,4.99,30,0,0

11,43,Female,White,Never,1,5.2,0,24.8,109,85,66,144,44,79,182,124,177,14.48,7.34,25.3,2,1

12,43,Female,Hobbit,Former,1,6.9,0,22.3,119,69,66,163,59,61,49,117,183,12.93,7.4,18.5,2,1

13,54,Female,White,Never,2,6.9,0,31.2,120,84,79,196,40,118,222,123,150,17.86,6.88,37.5,2,1

14,19,Male,White,Former,6,6.8,0,21.6,105,76,76,131,64,50,141,83,118,5.18,5.59,7.3,0,0

15,22,Male,Asian,Never,3,9.2,0,31.4,113,69,62,188,47,100,131,105,152,16.58,5.88,21.1,3,0