University of Zurich[UZH]

Big Data: Technology and Law

Predicting Diseases with Artificial Intelligence Using a Mobile-App

Implications of Law, Technology and Economy

Louise Cabot, Noah Chavannes, Martina Freund, Yves Rutishauser,
Charlène Zellweger

15.05.2020

Driven by technological advances in the medical field, the world has seen exploding health care costs. A first step towards a more efficient healthcare system is to reduce unnecessary visits at the doctor's. Advances in Artificial Intelligence allow us to detect possible diseases just by using an app. This reduces the amount of healthy people visiting doctors without any need and hence lowers the stress on the health care system. Artificial Intelligence requires a vast amount of health data which is protected by different laws and regulations, such as the GDPR. To get a better overview of how such an app could be implemented, we developed a website which allows the user to explore different configurations of the app. In every configuration, the implications on data protection law, economics and IT are shown. We especially highlight how these areas are interconnected. Additionally, we created an animation for each configuration that visually explains what happens to the user's data.

Access our website by using the following link: https://gdpr-ai.web.app

Table of Contents

1 Introduction

Recent technological changes have had a significant impact on the healthcare system. Digitalization enables our world to offer much more efficient treatment and diagnostic methods. Following this trend, our group came up with the idea of creating an app that could reduce unnecessary visits at the doctor's. This app would be able to detect diseases with the help of artificial intelligence algorithms. Thus, our project aims to cover and discuss various legal, technological and economic implications and seek to answer questions raised by the creation of this app.

Following this purpose, we decided to put together both a website and this paper. In both we will show important connections between legal, technical and economic aspects, which are all necessary for a successful potential launch of the app.

The website (https://gdpr-ai.web.app/) aims to give the visitors a user-friendly, understandable and mostly concise overview of the discussed matter. This paper aims to complement the website. Depending on the concerned field, the paper will give the reader more detailed information than the website. The main goal in creating the website was to give everyone access to the essence of our project, in a user-friendly form and with content that is easy to understand. The depth of the explanations is given by the complexity of the respective legal, technical or economic aspect.

Both on the website as in this paper the following topics will be discussed:

In the legal sections, you will get an insight into what the General Data Protection Regulation (GDPR) is. Based on this, we will show what rights the app users have with regard to their data but also what legal obligations our company has to observe with regards to data protection. In addition, we will discuss the legal implications, represented in the various possible functional models of the app. Creating an app raises legal questions in many different fields, such as liability, intellectual property, contract, etc. However, in the context of our project, the legal implications are only discussed with regards to data protection law.

In the technological sections, you will find how different configurations of an app require different deployment methods and different security aspects like encryption or access rights. Further we show what aspects in building an AI model need to be considered when choosing a deployment method.

In the economic sections, you will find a presentation of the advantages and disadvantages for our company, the users and society overall in each model. With regards to our company, aspects such as the use of data to improve services or the potentially highly profitable sale of data are discussed. For consumers, the focus is on the trade-off between data protection and data access. And for society overall, further external effects associated with the collection and distribution of data are presented.

Furthermore, this project should be seen as a pre-project to a more elaborated app. To make our statement clear and understandable, the app model of which we are discussing the legal

and technological implication is only able to detect one illness: tonsillitis. However, in the economic sections we decided to present the economic implications of the app as it would be in its final form: able to detect numerous different illnesses. This would be a more profitable and comprehensive approach, since an app that only detects tonsillitis would have very limited economic value.

Considering the possibility of a more elaborated version of the app, we think this project is of broad social interest. We think that such an app could revolutionise the way healthcare is traditionally administered in the European Union. Varying in its type of possible setting (offline, semi-online, online) our app could be very helpful to regions with restricted access to doctors but also to regions with low Internet coverage.

2 Presentation of the Legal Aspects

This first chapter of this paper endeavours to discuss the relevant aspects of the General Data Protection Regulation[1] ('the GDPR') and link them to our application in a concise manner.

The GDPR repeals the previous 1995 Data Protection Directive[2]. This new legislation takes the form of a Regulation and is directly binding on all EU and European Economic Area Member States. It is relatively recent and became enforceable as of May 2018. The GDPR aims at 'harmonising data protection law across the EU'[3] and affords individuals heightened 'control over their own personal data'[4].

We will address the legal questions raised by the creation of a health app in relation to the GDPR such as the extent of its applicability, collected data, accountability, consent, security and sale of data. These keys issues will be concisely discussed below, while a detailed user-friendly but general presentation of the GDPR's legal implications for our app is to be found in the legal section of the artefact.

Article 2(1) of the GDPR defines the material scope of the Regulation, which 'applies to the processing of personal data wholly or partly by automated means'[5]. We are thus directly concerned by the GDPR. Our fictional application is based in the European Union and is aimed at EU individuals. [6] The app's processing activities are related to the offering of a service, which aims to provide medical predictions.

[1] "REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL." EUR-Lex, 2016, eur-lex.europa.eu/eli/reg/2016/679/oj. (Last accessed 9.05.2020).
[2] Directive 95/46/EC.
[3] Christopher Kuner, Dan Jerker *et al*. 'The GDPR as a chance to break down borders' 7(4) *International Data Privacy* Law (2017) p 231.
[4] OJ L 119/2, 4.5.2016 (7).
[5] Article 2(1)
[6] It is worth noting that Switzerland has revised the Swiss Federal Act on Data Protection in order to implement the GDPR.

2.1 Categories of Collected Personal Data

Article 4(1) provides that personal data relates to any information that allows for direct and indirect identification of the data subjects, including 'an identifier such as a name, an identification number, location data, and online identifier (…)'[7]. In both the online and semi-online models of our health app requiring registration, we collect the password and email address (from the creation of an account), profile information and user content (photos) posted to the service.

An alternative for the user is to choose not to register and be able to use the app without direct identification. He will only need to provide a photo of his throat to receive a medical prediction. Hence, this alternative is more attractive to the user as he gives away significantly less data and thus will suffer limited consequences if there is a data breach.

The GDPR does not apply to our offline model. Indeed, the user is still required to provide a photo of his/her throat that will only be used to generate a medical prediction and afterwards, will be immediately deleted. No storage or processing activity takes place after being given a medical prediction.

Article 9 GDPR provides for a special category of personal data being collected: sensitive personal data. This concerns us as we also process 'data concerning health'[8] which is essential to fulfil the purpose of our app. Therefore, the images provided by the user are of a higher risk.

2.2 Accountability

Secondly, accountability for data protection is attributed to data controllers and data processors. Our company collects personal information and photos, determines the purpose of the processing and processes the data: our company is the data controller and does not resort to the services of a third-party processor.

It is possible for another actor to assume responsibility: the data processor who 'processes personal data on behalf of the controller'[9]. An obligatory, legally binding written contract governs the controller-processor relationship, also known as the Data Processing Agreement. *Article 28(3)* 'sets forth its minimum requirements'[10] to ensure that adequate provisions in the DPA can protect the user. In our models, we have decided not send data to a third party processor. Hence our company processes personal data without falling under the 'data processor' definition of the GDPR.

[7] Article 4(1) GDPR.
[8] Article 9 GDPR.
[9] Article 4(8) GDPR.
[10] *GDPR Processing*, Intersoft Consulting, gdpr-info.eu/issues/processing/. (Last accessed 13.05.2020).

2.3 Lawful Reasons for Data Processing and Consent

There are six[11] different legal basis possible to process data under *Article 6* GDPR. This justification informs users that we operate lawfully. In the context of our app, consent is the most appropriate legal basis to lawfully process data. The other five legal bases can, in certain circumstances, be more appropriate for certain apps.

Consent[12] is the legal basis we use to legally gain access to user data and process it. The standard for consent is high and requires users to genuinely understand the implications of giving their consent. Users of the app are required to make a genuine, informed choice to opt-in and agree to the processing of their data. Pre-ticked boxes are not allowed. For each of our models, consent is requested as the user starts to use our services. Our privacy policy provides in detail what the user has consented to.[13]

Article 9(1) prohibits the processing of sensitive personal data unless justified by one of the exceptions listed under *Article 9(2)*[14]. 'Explicit' consent is one of the exceptions (compared to 'informed' consent required when processing personal data). This is the legal exception our company relies on the process sensitive personal data.

2.4 Transfers and Sale of Data

In certain of our models we sell the data collected to thirds parties located in the EU, such as to marketing and business partners. This shall only take place on the basis of a legal permission. In our situation only consent is a valid lawful reason.

Additionally, transfers of personal data outside the EU/EEA are very common, as they further international trade and cooperation between countries. However, restrictions do apply. *Article 45(1)* of the GDPR provides that the European Commission is tasked to determine whether there is an adequate level of data protection in the third country: this is called an adequacy decision. Such a decision makes it possible for personal data to 'flow from the EU (and Norway, Liechtenstein and Iceland) to that third country without any further safeguard being necessary'[15].

Amongst others, the United States, Switzerland, Japan and New Zealand have been recognised as providing sufficient protection to personal data coming from the EU.

[11] Consent; contract; legal obligation; vital interests; public task; and legitimate interests.
[12] Article 6(1)(a) GDPR.
[13] Article 7(2) imposes as a condition of consent the obligation to provide a privacy policy.
[14] In addition to fulfilling one of the lawful bases provided for in Article 6.
[15] European Commission. *Adequacy Decisions*. Justice and Consumers, 2019, ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en. (Last accessed 12.05.2020).

If the third country has not qualified for an 'adequacy decision', *Article 46* makes data transfers possible only if 'appropriate safeguards' are implemented.

2.5 Security

To minimise the chance of a data breaches, it is highly advisable that any data controller carry out a data protection impact assessment ('DPIA') to best determine the approach to be taken. In our situation such an assessment would be compulsory.

In order to securely store and process the personal data obtained from the user, *Article 5(1)(f)* recognises the need for controllers to follow the 'integrity and confidentiality' principle. As the harm caused by the loss of the user's personal data could include identity fraud and targeting of individuals by fraudsters, we aim to 'implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk'[16]. After analysing the risks inherent to our data, we have endeavoured to use encryption and pseudonymisation measures. Encryption is just one of the appropriate measures possible to safeguard the security of personal data being stored.[17]

Additionally, any of the data stored in our services abides by the 'storage limitation' principle found in *Article 5(1)(e)* which requests that any personal data we no longer need is to erase.

In addition, *Article 33* requires us to notify the relevant Data Protection Authority when a data breach occurs. Documenting any breaches is highly important and will help us stay accountable to the users of our app.

2.6 Protected Individual Rights

The GDPR provides data subjects with several rights which we have sought to respect during the collection and use of personal data.

When starting to use our app, a Privacy Policy will be available to help users fully comprehend the extent of their rights. This requirement of providing adequate notice and respecting the right to be informed is enshrined in *Article 13*.

According to *Article 21*, users have, at any time during the use of the app, the right to object to their data being processed. This right is of high importance to users when we engage in the selling of their personal data to marketing partners.

Additionally, the right to erasure of personal data is provided for in the GDPR under *Article 17* and covers personal data that has been stored in the app's backup systems. This is a far-reaching right which recognises the user's power and possibilities of control. However, this

[16] Article 32(1) GDPR.
[17] See Article 32 for further appropriate technical measures.

right is not absolute. For instance, if a company's main purpose is the performance of a task in the public interest, the right to erasure will not apply and it will be able to continue processing personal data. The extent of the right to erasure all depends on the purpose of the data processing. As we solely rely on obtaining consent from users as 'lawful basis for holding the data'[18], they are able to withdraw consent at any time. We would be obliged to delete any personal data collected and stored. This has a negative impact on our health app's functioning as its aim is to provide better, accurate predictions which can only be done when the photos of our users enable us to train our system.

Lastly, when a data breach 'affects the confidentiality, integrity or availability of personal data'[19], controllers are required to examine the risks 'to people's rights and freedoms'[20]. In cases of severe breaches, controllers are obliged to report the designated Data Protection Authority of the EU country and may also be required to notify the affected user. *Article 82* provides users for the right to compensation if it transpires that we have failed to comply with all GDPR requirements. For example, if we fail to notify a user of a breach of their personal data and it has 'suffered material or non-material damage as a result of an infringement'[21], then we are liable for immediate compensation.

2.7 Conclusion

To conclude, the GDPR is an effective tool to protect individual's rights and has significantly improved trust, transparency and better decision-making. The clearly defined rules have provided legal clarity which benefits both individuals and businesses.

Complying with the GDPR as an app developer requires many obligations and steps to follow which will be further developed in the legal section of the artefact.

However, much of the GDPR provisions leave scope for interpretation and full compliance has appeared to be onerous for businesses. It has been criticised, *inter alia,* as imposing a set of new requirements on businesses and several unnecessary administrative burdens.

This chapter of the paper has sought to give a comprehensive overview of the GDPR and to relate it to our health app.

[18] *Right to Erasure*, ICO, ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-erasure/?q=privacy notice . (Last accessed 13.05.2020).
[19] *Personal Data Breaches*. ICO, ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/personal-data-breaches/?q=liability . (Last accessed 04.05.2020).
[20] Ibid.
[21] Article 82(1) GDPR.

3 Presentation of the Technical Aspects

In this chapter we will discuss the AI model and the security aspects of the application. The AI model represents the core of the application which enables generating medical predictions based on the user's picture. Furthermore, securing the user's sensitive health care data is important and hence we outline the most important security measures.

3.1 Model Architecture and Training Model

In this subchapter, we will introduce the idea behind a model, which detects whether a person is affected by tonsillitis. First, the general concept behind deep learning models will be explained. Then, we will describe how the pictures are transformed such that they can be processed by the model. Furthermore, we will present the model's architecture and explain its key features. Additionally, we will highlight how the model is evaluated. Lastly, we will explain how to the model could be extended to detect more throat diseases.

3.1.1 General Concept Behind Deep Learning Models

Deep learning algorithms [22] are extremely popular among image classification. Neural networks, which are a part of deep learning, try to imitate the human brain using mathematical and statistical models. Especially convolutional neural networks (CNNs)[23] pose a way to detect patterns in images and are then able to learn how to classify them correctly. CNNs belong to supervised learning techniques, i.e. a training dataset of images and their corresponding labels. Labels in this context refer to the classification whether the throat in the image is affected by tonsillitis. The training set in our specific task consists of images of healthy throats and throats with tonsillitis. It is important to note that the training images have to be correctly labelled, otherwise the CNN is not able to learn from them.

3.1.2 Preprocess Data

In this part, we discuss the pre-processing of images. For a better illustration, we will use a picture to exemplify the procedure. In more traditional approaches, the pre-processing of images includes manually defining appropriate filters that are used for feature extraction. Since convolutional neural networks largely replace these filters[23], little effort will be taken to pre-process the images. For simplicity and data reduction, the pictures will first be grey scaled[24].

[22] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
[23] LeCun, Yann, et al. "Object recognition with gradient-based learning." *Shape, contour and grouping in computer vision*. Springer, Berlin, Heidelberg, 1999. 319-345.
[24] Ĉadík, M. "Perceptual evaluation of color-to-grayscale image conversions." *Computer Graphics Forum*. Vol. 27. No. 7. Oxford, UK: Blackwell Publishing Ltd, 2008.

An important constraint for CNNs poses the fact that the images need to be of a unified dimension, all the pictures need to be resized to a specified height and width. Further pre-processing steps need to be done in order to train the network with the images (e.g. normalization).

3.1.3 Convolutional Neural Network (Model architecture)

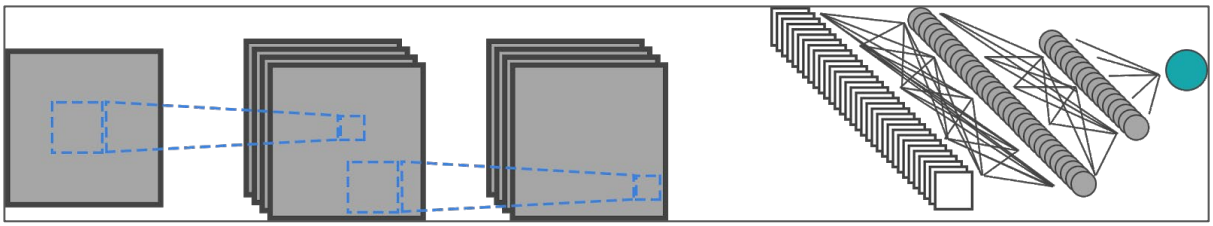

Figure 1: CNN with 2 filters and a densely connected neural network

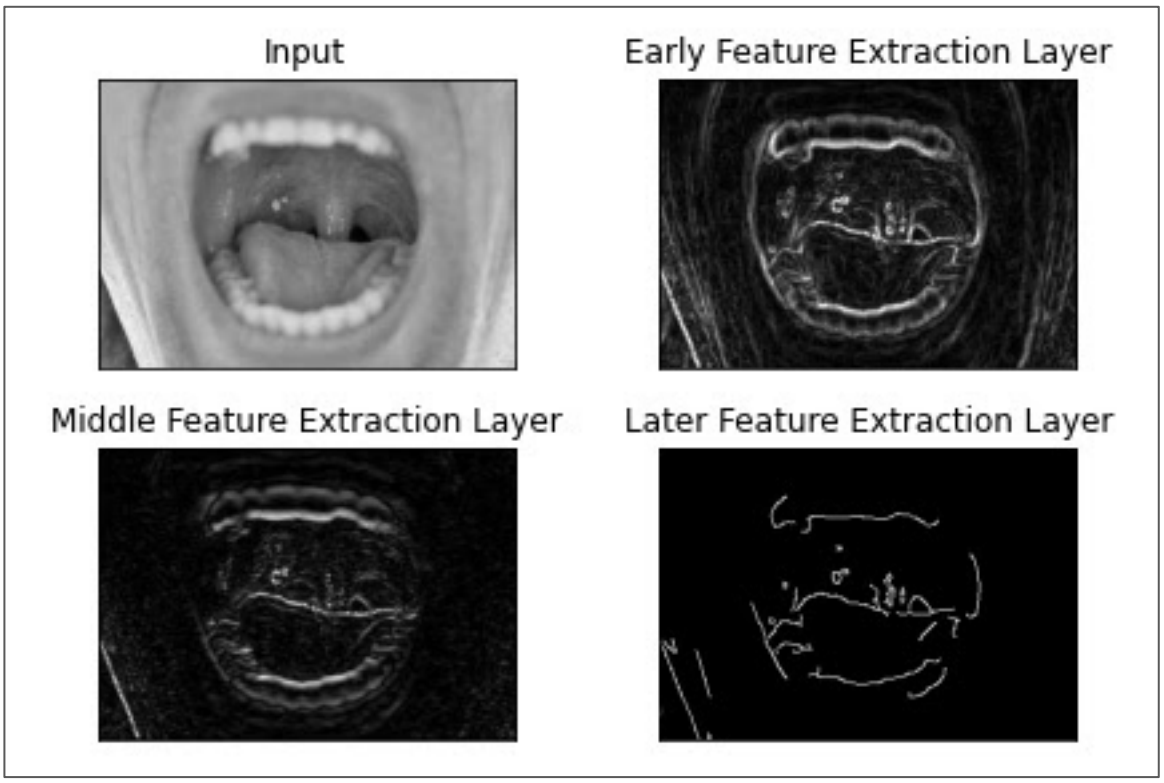

Figure 2: Output of feature extraction in different layers

Convolutional neural networks[25] allow to input the normalized raw image and retrieve a decision as the output. Figure 1 visualizes a simplified architecture of such a model. The different stages of the model are explained from left to right: First, the model reads the image. The grey-scaled image is stored in a matrix, one cell for each pixel. Then, filters slide over the

[25] "Convolutional Neural Networks." *CS231n Convolutional Neural Networks for Visual Recognition*, cs231n.github.io/convolutional-networks/. (Last accessed 05.05.2020).

image to detect local patterns and pass this information to the first feature extraction layer. The first feature extraction layer is able to detect edges and lines while later layers detect more high-level features (such as the texture of the teeth). A visualization of this explanation can be seen in figure 2. The first feature extraction layer is similar to the raw input image. The further away a feature layer is from the input image, the more abstract the picture becomes. The later feature extraction layer visualizes the pictures just with white lines. If you look precisely, you can still see the tonsillitis blister marked on the last layer. Hence, the neural network was able to extract an important part of the image.

The filter size and specific amount of feature extraction layers are hard to determine beforehand and are therefore hyperparameters that can be set and adjusted by the developer during training the model.

The last feature extraction layer is flattened from a 3-dimensional matrix to a 1-dimensional vector and fed into a fully connected neural network. The last layer outputs the probability distribution. Since in figure 1 only one output node (blue) is present, this model suggests a binary classification task. In our use case, this refers to whether a user's picture of the throat results in the disease tonsillitis or not.

3.1.4 Training the Model

AI models require to be trained which refers to the process of learning how to be effective in extracting information and calculating predictions. Training CNNs often requires a large dataset. Techniques such as data augmentation or upsampling are used to enlarge sparse datasets[26]. The model is trained using a balanced dataset (roughly 50% with the disease and 50% without the disease). The images are first pre-processed and then fed into the neural network. The weights of the network are adjusted during training through backpropagation. Backpropagation is the application of mathematical algorithms that allow the calculations of derivatives, which are needed to optimize the model, without having to recalculate every step along the way. Therefore, backpropagation reduces the computational cost of training[27].

3.1.5 Model Evaluation

The high amount of hyperparameters available due to the architecture of the model and the many possibilities to pre-process an image gives us a lot of flexibility to tweak our model. Hence, we will train our model using different hyperparameters and different pre-processing techniques such as blurring unnecessary noise (e.g. background) or using coloured pictures.

[26] Inoue, Hiroshi. "Data augmentation by pairing samples for images classification." *arXiv preprint arXiv:*1801.02929 (2018).
[27] Hecht-Nielsen, Robert. "Theory of the backpropagation neural network." *Neural networks for perception.* Academic Press, 1992. 65-93.

In our use case, we are not only interested in increasing accuracy (total amount predicted correct by the model / total amount of data points) but we also focus on the statistical measure recall (True positive / (True positive + False negative))[28].

Consider 10'000 pictures in our database where 5'000 pictures contain the disease. Our classifier should be able to detect a high number of the pictures containing the disease. Hence, we want to minimize the case where a person actually has tonsillitis, but the model wrongly predicts that the person is healthy. Based on the recall we will adjust the hyperparameters and pre-processing strategy.

3.1.6 Future Extension of the Model's Capabilities

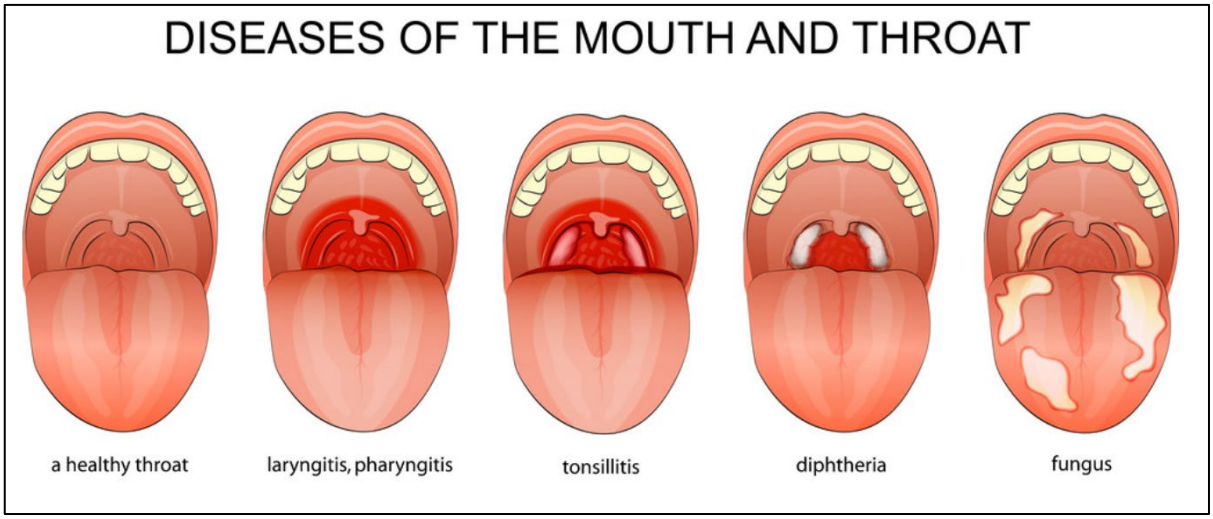

Figure 3: Diseases of the throat *(Source: https://www.vectorstock.com/7612153)*

The model explained in this subchapter is a very simplified version of the models used in an actual production setting. In a realistic setting, the user would first be asked which part of the body he/she wants to analyse. Second, the model would be able to detect one of many different diseases. For example, the model would be able to predict that a person has diphtheria but is not healthy and does not have tonsillitis. In this version of the model, the pictures would probably not be grey scaled, since the colour could contain important information on the disease.

[28] Powers, David Martin. "*Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.*" (2011).

3.2 Infrastructure

In this chapter we will discuss the security measures that protect the users' sensitive data. This includes the security during transfer and storage. Furthermore, we will outline access restrictions.

3.2.1 Security During Transfer

In order to protect the sensitive data that is sent from a user's device to the company's server, the Transport Layer Security protocol (TLS)[29], also known as SSL[30], is followed. SSL encrypts the data during the transfer. This does not guarantee the safety of the data once it has arrived on the company's servers. In order to ensure this, server-side encryption as well as state of the art security systems have to be in place.

3.2.2 Encryption

When handling sensitive data, it is important to have measures in place to protect the data during transfer between devices as well as whilst stored in databases. Data breaches without having such measures in place can have a big economic impact on the company responsible for the breach as well as for the people affected.

That's why we suggest every company have some encryption techniques in use, such that they comply with GDPR and reduce the economic risks in case of a system breach.

Encryption is a way of translating a readable text (e.g. the users name) into a ciphertext that cannot be interpreted anymore. This translation is done by applying mathematical algorithms to the initial text. Encryption algorithms use a set of keys which are needed for encryption and decryption. Furthermore, there are symmetrical algorithms which use the same key for encryption as well as for decryption and asymmetrical keys which use different keys. For our application we propose the use of a symmetric algorithm, as we want to be able to revert the encryption when needed. A symmetrical encryption algorithm that would provide sufficient security during storage of the data would be the Advanced Encryption Standard (AES)[31] with a 256-bit (32 character) long key.

If the AES algorithm is applied on the text Username with the 256-bit key *super_secret_special_company_key* the resulting ciphertext would be *45at7ovrFcmmHp35LmXvJw==* (base64 encoded for readability). This means that in the case

[29] "The Transport Layer Security (TLS) Protocol Version 1.3." *IETF Tools*, tools.ietf.org/html/rfc8446. (Last accessed 03.05.2020)
[30] "How Does SSL Work?" *How Does SSL Work? | Entrust Datacard*, www.entrustdatacard.com/pages/ssl. (Last accessed: 10.05.2020)
[31] Daemen, Joan, and Vincent Rijmen. *The design of Rijndael: AES-the advanced encryption standard.* Springer Science & Business Media, 2013.

of a data breach only the not human readable ciphertext would be exposed. If the attacker would want to revert the encryption without knowing the key, he/she would have to $1.1 * 10^{77}$ different combinations, which is with today's computing power impossible. But AES 256-bit it only as secure as the key is. If the encryption key is exposed, a cybercriminal would also have access to all the data in human readable form.

Besides securing data, encryption has also a further benefit. It can be used for pseudonymisation[32]. Pseudonymisation is a technique where identity revealing attributes are replaced by another value, that is not related to the subject anymore.

Compared to anonymisation, where all identity revealing attributes are thrown away, pseudonymisation keeps the attributes but replaces them with pseudonyms. An important trade of pseudonymisation is that data from different data sets can still be related to each other, due to the fact that if the same pseudonymisation technique is applied, the generated pseudonym is the same. This can be very important in research as well for selling data without compromising the user's identity. Pseudonymisation in our case can be done by applying the AES algorithm on relevant attributes or even the whole database.

3.2.3 Authentication and Authorization

Authentication is the process of a user having to prove the identity to the server. When the user registers, authentication is given with the user's email and password. The user can then further secure his/her account by using two-factor-authentication. If two-factor-authentication is enabled and after successfully entering username and password, the user has to provide a second authentication factor (e.g. SMS token or time-based token). This ensures that the person trying to log-in not only knows a secret that only the real user would know, but that he/she is also in possession of an item (e.g. Phone) only the real user would have. This allows for better protection of the access to the sensitive healthcare data.

In case the users do not have to register to use the app, a device-based, hashed ID for each user will be created. The purpose of the hashed ID for each user is to grant him/her access rights to his/her data during authorization without the need of providing any personal data.

After a user is successfully authenticated, he also needs to be authorized to be able to request and send data. The authorization will be handled by the company's servers with the use of token-based endpoint authorization (e.g. OAuth2[33]). With proper implementation of such a protocol, we can make sure that every user can only access his/her data. This methodology works for cases where a user has to register but also when only a hashed device ID is available. Furthermore, this enables automating the process of deleting user data as soon as it is requested.

[32] Tinabo, Rose, Fred Mtenzi, and Brendan O'Shea. "Anonymisation vs. Pseudonymisation: Which one is most useful for both privacy protection and usefulness of e-healthcare data." *2009 International Conference for Internet Technology and Secured Transactions,* (ICITST). IEEE, 2009.
[33] *OAuth2*. https://oauth.net/2/ (Last accessed 11.05.2020).

It's important to note that the data is deleted immediately, but the model that has already been trained with this data is still active. The data will only be truly gone, after the model is retrained.

4 Presentation of the Economic Aspects

In this section we will show how the business model in terms of data collection and data use could look like for our company when commercializing the app. Companies use data and the benefits derived from it to differentiate themselves from their competitors. In this context, data is also referred to as the new oil[34] and thus seen by companies as an asset, currency or for monetization purposes. This will enable us to use the collected data to make (additional) profits. In the discussion of the different business models, we will therefore show that our company has different ways to generate money: Either by selling and collecting data and/or by charging subscriber fees.

4.1 Data as an Asset

A data set can be owned and controlled, is exchangeable for money and has the potential to generate future economic benefits[35]. Therefore, companies use data in their decision-making processes, for compliance purposes or for improvements of the customer service. The right use of business data enables a firm to fully leverage the data's potential for competitive advantages. The more data our company could aggregate and process, the higher the barriers for new market entrants[36]. Without having a large user base, it is hard to generate big datasets at first.

Ideally, our company starts collecting as much data as possible. Though, there are certain characteristics that make data more or less valuable. The most valuable datasets are hard to replicate for other companies without having the same user base. Further the company should pay attention on keeping the data clean, accurate and up to date. The more useful applications and the more positive network effects they imply, the higher the additional value of the data[37].

[34] Fischer, P. "Editorial. Big Data – Der Wertvolle Rohstoff Der Informationsgesellschaft." *Die Volkswirtschaft*, vol. 87, no. 5, 2014, p. 3.
Schmarzo, B., and M. Sidaoui. *Applying Economic Concepts to Big Data to Determine the Financial Value of the Organization's Data and Analytics, and Understanding the Ramifications on the Organizations' Financial Statements and It Operations and Business Strategies*. 2017, infocus.dellemc.com/wp-content/uploads/2017/04/USF_The_Economics_of_Data_and_Analytics-Final3.pdf. (Last accessed 13.04.2020).
[35] Laney, D. "3D Data Management: Controlling Data Volume, Velocity, and Variety. ." *META Group Inc. APPLICATION DELIVERY STRATEGIES*, 2001. blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (Last accessed 13.04.2020).
Woodie, A. *How Do You Value Information*. Datanami, 2016, www.datanami.com/2016/09/15/how-do-you-value-information/. (Last accessed 13.04.2020).
[36] Southekal, P. H. "Data for Business Performance: The Goal-Question-Metric (GQM) Model to Transform Business Data into an Enterprise Asset." *Technics Publications, LLC*, 2017.
[37] Polovets, L. "The Value of Data." *Coding VC*, 2015, www.codingvc.com/the-value-of-data-part-1-using-data-as-a-competitive-advantage. (Last accessed 02.04.2020).

Therefore, our data should be collected as much as possible, as early as possible, as raw as possible and with as many connections to other datasets as possible. By evaluating how and what data should be collected, value diminishing aspects should also be taken into account. The value of the data decreases over time, either because the data become older or because the marginal utility becomes less and less. This applies, for example, to data used for AI training purposes. After a certain amount of input variables, the value of an additional input variable for making better predictions starts decreasing. Additionally, the value of data depends on how easy it is to replicate them in a different way[38]. This applies to healthcare data in general since it is hard to receive accurate data simply from data exhaust. Thus, third parties will pay higher amounts for our data since it is more difficult to generate from scratch.

Data can be distinguished into three types: self-reported data, digital exhaust and profiling data[39]. Self-reported data refers to any information that is given voluntarily by the user, for example, through surveys or the entry of email and home addresses. In our case it would be the uploaded picture of the disease's symptoms or the email addresses in the case of registration. Digital exhaust is mostly more detailed and results from the users' behaviour (purchase decisions, like rates, time or location measurements). The most precise data arises when different sets of self-reported data and digital exhaust are combined in order to establish data profiles of individual users. This enables firms to make predictions about the users' interests and behaviours[40]. For example, if our company were to sell directly identifiable information to third parties, consumers could be reached with advertising specifically targeted at the before identified diseases. However, this brings us to the most important trade-off problem: The more personalized and thus more valuable the data is, the more sensitive it is for consumers.

"Without customers, we can have no business." This motto requires us to focus not only on the value of possible data sets but also on the preferences of our customers. According to a study that evaluated how much people value their personal data, in Germany and Great Britain, health care data is the data most valued. Germans, for instance, estimate the personal value of their health history at more than 180 USD[41]. Furthermore, Morey et al. found that 97% of the surveyed people expressed concerns that businesses might misuse their data. This means that a business model that includes healthcare data like our app should be especially careful in balancing the trade-off between making profit with collected data and protecting consumers' privacy.

[38] Polovets, L. "The Value of Data." *Coding VC*, 2015, www.codingvc.com/the-value-of-data-part-1-using-data-as-a-competitive-advantage. (Last accessed 02.04.2020).
[39] Morey, T., et al. "Customer Data: Designing for Transparency and Trust." *Harvard Business Review*, vol. 2, 2015, pp. 96–105., https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust . (Last accessed 13.04.2020).
[40] Ibid.
[41] Ibid.

4.2 The Business Models

Related to this above-mentioned trade-off, the firm has to balance the advantages of collecting and selling data against its disadvantages. In this section, we will present three major business models varying in the collection and usage of data.

4.2.1 No Collection of Data

In order to completely avoid this trade-off, our company can decide to simply not collect any data, which would be the case in the offline model. The consumer's privacy is fully protected, and the firm does not need to check compliance with the GDPR. However, this implies that our firm can also not derive any value from the data. Compared to other firms that collect and maybe even sell their data this can be a huge competitive disadvantage[42]: customer services cannot be improved, the AI model cannot be further trained and switching costs for consumers can also not be improved through personalizing the content and statistics of the app.

4.2.2 Data Collection Only

Another approach of the company could also imply the collection but not the sale of data. Here the firm can still derive value from the data. Through the collected data, the customer service can be measured and improved, the application can be improved, app marketing can be fostered, and the AI's algorithm can be trained[43]. Nevertheless, the perceived privacy of consumers remains mostly protected. Additionally, the risk of underestimating possible (legal) consequences when exposing the data to third parties can be eliminated[44]. However, the company must be aware of possible data breaches and the associated costs. Risk Based Security[45] found that 7,098 data breaches were reported worldwide in 2019, resulting in the exposure of 15.1 billion records. In this model, also customers' confidence might be harder to

[42] Acquisti, A. "The Economics of Personal Data and the Economics of Privacy." 2010, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1017.7187&rep=rep1&type=pdf . (Last accessed 30.04.2020).
[43] Ibid.
Faktor, S. *Never Sell Data*. Forbes, 2014, www.forbes.com/sites/stevefaktor/2014/03/25/never-sell-data/#5cb41d39259a. (Last accessed 23.04.2020).
[44] Groves, P., et al. "The 'Big Data' Revolution in Healthcare. Accelerating Value and Innovation." *McKinsey & Company Report*, 2013, https://www.ghdonline.org/uploads/Big_Data_Revolution_in_health_care_2013_McKinsey_Report.pdf. (Last accessed 08.05.2020).
[45] Risk Based Security. "Data Breach Quick View." *Data Breach Quick View Report*, 2019, pages.riskbasedsecurity.com/2019-year-end-data-breach-quickview-report. (Last accessed 01.05.2020).

gain[46]. Wottrich et al.[47] found that when downloading an app, customers indeed engage in a privacy trade-off. When deciding how much data they want to share, users consider several aspects such as the value of the app, the expected consequences of their data access permission, the apps perceived intrusiveness and the given permission justification of the app. Customers are much more willing to share their data, the more valuable the services are, the more popular the app is and the less intrusive data is requested. Furthermore, customers appreciate it if the use of their data is explained very transparently[48]. A disadvantage for the customers still might arise if our company is not fully able to cover all the costs when not selling the data. Then the users could be asked for a one-time download fee or a recurring subscription fee.

4.2.3 Collection and Sale of Data

As a third approach the company can also decide to collect and sell its data. This offers the company the most opportunities to benefit from the data collected, especially if the collected data is directly identifiable data. Spiekermann et al.[49] found that 89% of identified profiles consist of highly rated data, while in anonymous profiles 78% are rated low and only 22% have a medium value for companies. So, if the company sells data that is directly identifiable, it can charge much higher prices than for not directly identifiable data. The OECD[50] estimated that an address is worth 0.5 USD whereas a social security number already accounts for up to 8 USD. In regard to the privacy trade-off, the company should ensure a high degree of transparency about the use and distribution of the data towards its customers. Spears[51] found that online customers tend to perceive a higher level of privacy security when they are informed about a firm's information practices. Customers highly value having insights into what the firm is doing with their personal data[52].

[46] Malhotra, N., et al. “Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model.” *Information Systems Research*, vol. 15, no. 4, 2004, pp. 336–355., https://www.jstor.org/stable/pdf/23015787.pdf?refreqid=excelsior:d0b7395ff7fd835c563e4bbb8f148842 . (Last accessed 01.05.2020).
[47] Wottrich, V., et al. “The Privacy Trade-off for Mobile App Downloads: The Roles of App Value, Intrusiveness, and Privacy Concerns.” *Decision Support Systems*, vol. 106, 2018, pp. 44–52., https://www.sciencedirect.com/science/article/pii/S0167923617302221 . (Last accessed 01.05.2020).
[48] Gu, J., et al. “Privacy Concerns for Mobile App Download: An Elaboration Likelihood Model Perspective.” Decision Support Systems, vol. 94, 2017, pp. 19–28.
[49] Spiekermann, S., et al. “User Agents in E-Commerce Environments: Industry vs. Consumer Perspectives on Data Exchange.” *Eder J., Missikoff M. (Eds) Advanced Information Systems Engineering. CAiSE 2003. Lecture Notes in Computer Science*, vol. 2681, 2003, https://link.springer.com/content/pdf/10.1007/3-540-45017-3_46.pdf . (Last accessed 30.04.2020).
[50] OECD. “Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value.” 2013, https://read.oecd-ilibrary.org/science-and-technology/exploring-the-economics-of-personal-data_5k486qtxldmq-en#page19. (Last accessed 08.05.2020).
[51] Spears, J. L. “The Effects of Notice versus Awareness: an Empirical Examination of an Online Consumer's Privacy Risk Treatment.” *Proceedings of the 46th Hawaii International Conference in System Sciences*, 2013, pp. 3229–3238.
[52] Malhotra, N., et al. “Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model.” *Information Systems Research*, vol. 15, no. 4, 2004, pp. 336–355.,

4.3 Key Insights from Business Models

To conclude the models discussed above, the company should find a more balanced way for its business model[53]. Since the app acts in the highly sensitive area of personal healthcare data, the company should not only look on the benefits of data collection and data sale but also on the shortcomings. In the case of mandatory disclosure of private health data to third parties, consumers will be strongly deterred from using the app. Thus, the firm could decide to just collect data but only using it for internal improvement purposes. The apps content and statistics can be personalized and like this the customer's switching costs will increase. Furthermore, the AI can be trained, and other customer improvements can be made by processing the data. Another advantage of not selling the data is the possibility of creating disadvantages for potential market entrants. The more data they had at their disposal, the easier it would be for them to start similar services.

Another possibility would be to give consumers the choice whether or not their data may be disclosed to third parties. In the event that no data is transferred, users would pay a small monthly fee to continue using the app and its services. This would provide transparency and allow users to retain control over the distribution of their data. Consumers who are not willing to pay the fee could remain with the basic account which would enable our company to sell their data. In this way, we would end up with two groups of consumers, one of whom would pay for the service with their data and the other with a subscription fee.

In total, our model could be included into a healthcare insurance. Customers who use the app before visiting the doctor would save on insurance costs. This could be awarded by insurance fee reductions[54]. Thus, the collection of directly identifiable data brings another advantage to our app users.

5 Future Work

In this work, we have discussed various legal, technical, and economic aspects of a health care application that is able to detect tonsillitis based on a user's image. In a realistic setting, such an app has to be able to detect a high variety of throat diseases. Furthermore, the app should include other models that are able to detect diseases in other parts of the body. These models

https://www.jstor.org/stable/pdf/23015787.pdf?refreqid=excelsior:d0b7395ff7fd835c563e4bbb8f148842 . (Last accessed 01.05.2020).

[53] Spiekermann, S., et al. "User Agents in E-Commerce Environments: Industry vs. Consumer Perspectives on Data Exchange." *Eder J., Missikoff M. (Eds) Advanced Information Systems Engineering. CAiSE 2003. Lecture Notes in Computer Science*, vol. 2681, 2003, https://link.springer.com/content/pdf/10.1007/3-540-45017-3_46.pdf . (Last accessed 30.04.2020).

[54] Jentzsch, N. *The Economics and Regulation of Financial Privacy. Contributions to Economics*. Physica-Verlag HD, 2006.

can be accessed by answering a questionnaire or tapping on an image of the body to indicate where the pain occurs.

In the chapter related to the aspect of law, we have emphasized on data protection. To further extend our work, the question of liability would need to be more thoroughly addressed, especially when the user has an illness, but our model is not able to detect it. For example, in the U.S., it is best practice to follow the U.S. Food and Drug Administration's (FDA) Guideline[55] when developing such an app in order to reduce liability when sued.

A further direction to extend our work could be a collaboration with a health care insurance provider. These institutions could use the app to reduce their costs by advising their clients to use the app before going to the doctor. However, this raises ethical dilemma because health care companies would be able to get an extensive insight into users' medical history.

6 Conclusion

We have created a website that visualizes the interaction between legal, technical, and economic aspects of a health care app that is understandable for everyone. The enclosed paper discussed these aspects in a more advanced manner.

All things considered, we have found that these aspects are deeply interconnected, especially when working with sensitive data. Decisions made when developing the app, have huge implications on legal matters and can lay the ground for a successful business campaign. On the other hand, the GDPR can restrict or complicate the app's internal technologies, e.g. obligation of encrypting the data for secure storage. Furthermore, the Regulation can have an impact on how to design a well-balanced and sustainable business strategy.

Additionally, the configurator visualizes that the app can be implemented in different ways, having various advantages and disadvantages. In most cases the online configuration, where a user has to register, makes the most sense. On the other hand, a semi-online configuration could be useful in a setting where the user doesn't have to rely on an active internet connection, e.g. when he/she wants to use the app in a third world country. The distribution of the data to other companies could also be left to the user's responsibility. Hence, a user could utilize the service for 'free' by paying with his/her data or by paying a small subscription fee but restricting the option of data being further distributed.

To conclude, this interdisciplinary project gave us a chance to collaborate beyond everyone's field. It gave us an insight into how the development of such an app might look like in practice.

[55] "Policy for Device Software Functions and Mobile Medical Applications" *Guidance for Industry and Food and Drug Administration Staff,* https://www.fda.gov/media/80958/download. (Last accessed: 10.05.2020).