*Your name here*

*Your report title here*

Table of Contents

[1. Introduction 1](#_Toc99955248)

[2. Background 1](#_Toc99955249)

[3. Problem statement 1](#_Toc99955250)

[4. Project objectives 2](#_Toc99955251)

[5. Data warehouse details 2](#_Toc99955252)

[6. Interesting insights 2](#_Toc99955253)

[7. Data pre-processing 2](#_Toc99955254)

[8. Measuring performance 2](#_Toc99955255)

[9. Algorithms applied 3](#_Toc99955256)

[9.1 Algorithm 1 3](#_Toc99955257)

[9.2 Algorithm 2 3](#_Toc99955258)

[9.3 Algorithm 3 3](#_Toc99955259)

[9.4 Comparison 3](#_Toc99955260)

[10. Discussion 3](#_Toc99955261)

[11. Conclusion 3](#_Toc99955262)

[References 4](#_Toc99955263)

1. **Introduction**

Skin diseases affect almost 900 million people in the world at any time (World Health Organization, 2018). The diseases can impact an individual’s quality of life and health. In this report we will present three machine learning models that can classify eight different skin diseases. It will include the performance and accuracy of each model, as well as comparisons.

By making a machine learning model that can classify skin disease fast and accurately, it can be a tool for doctors to diagnose skin diseases. Furthermore, it can increase efficiency and reduce costs in the health care system.

1. **Background**

The healthcare industry is one of the biggest industries in the world (Wikipedia,2023). This is one of the reasons why we chose to base our project on this domain. The possibilities machine learning can have in the healthcare industry is monumental and can be used as a tool to help improve both the quality of care and efficiency.

According to (Džakula et al., 2022) the shortage of healthcare workers is a worldwide problem. Additionally, the health care industry is estimated to need 80 million more workers by 2030 to meet demand (Deloitte,2023). This highlights the need for tools to make the system more efficient, but also make the machine learning model that we created a tool for healthcare workers by maybe making their job easier. Furthermore, it can also help the patients who are affected with skin diseases get diagnosed and get treatment faster.

**2.1** **Skin diseases**

A skin disease is a condition that affects the skin. The skin disease can cause inflammation, rashes, itchiness and/or other skin changes (Cleveland Clinic, 2021). For example, cellulitis, which is one skin disease we are training the machine learning model to classify, it can cause redness, swelling, warmth, and pain in the affected area. There are also different causes to skin diseases, e.g., bacteria, fungal infections, or viruses (Higuera, 2023). The skin diseases we are learning the machine learning model to identify can be caused by bacteria, fungal, parasites and viruses. The cause and the names of the skin diseases we are classifying is:

1. Bacterial Infections- cellulitis
2. Bacterial Infections- impetigo
3. Fungal Infections - athlete -foot
4. Fungal Infections - nail-fungus
5. Fungal Infections - ringworm
6. Parasitic Infections - cutaneous-larva-migrans
7. Viral skin infections - chickenpox
8. Viral skin infections – shingles

So, there is four different causes to the skin diseases that the machine learning model is going to classify. Some of these skin diseases are common, other ones are rarer. For example, cellulatis is a common bacterial skin infection (Cellulitis: All you need to know 2022). Other ones are less common, for example, cutaneous larva migrans, because this is most common in tropical environments (Dinulos, 2023). Unfamiliar skin diseases can be a challenge for doctors due to their limited exposure to these cases. Next section we will look at the diagnosing and treatment of skin disease.

* 1. **Diagnosing and treatment**

In a study done the results were that nondermatologists perform poorly in both diagnosing and treatment of skin diseases (Federman et al., 2004). The most common way for a doctor to diagnose skin disease is by visually examining the skin (Cleveland Clinic, 2021). Which is why a machine learning model that is trained on images of skin disease can be an effective tool. Some of the skin diseases we are training the machine learning model to classify is a little similar, e.g., shingles and chickenpox. Because of this it can be difficult for a doctor who is not specialized in skin disease to diagnose the skin disease accurately. With the machine learning model, we want to make a model that is enough accurate to distinguish between these two. There are several ways to treat a skin disease. Some of the skin diseases may also go away by its own. Some common treatments of skin diseases include creams, lotions, oral medication, light therapy, and surgery (Cleveland Clinic, 2021).

There are several challenges associated with diagnosing and treating skin diseases. One challenge is that there are many different types of skin diseases, and they can sometimes be difficult to distinguish from each other, which is mentioned above. Another challenge is that some skin diseases, such as shingles, can be difficult to manage and treat (Johns Hopkins Medicine, 2023). Furthermore, a machine learning model can maybe faster diagnose a skin disease, which can lead to faster treatment. If we take shingles as an example, it is important to start treatment early with this skin disease because some medication is not as effective if the rash has been present for 2-3 days (Johns Hopkins Medicine, 2023).

* 1. **Machine learning**

Machine learning has the potential to be a powerful tool to the diagnosis and treatment of skin diseases. Machine learning models can be trained on large datasets of images of skin diseases to learn to identify patterns that are associated with different skin diseases. This can help doctors to diagnose skin diseases more quickly and accurately than traditional methods. We hope that trough the machine learning model that we have made, we will be able to achieve this.

In the future, machine learning could play an even greater role in dermatology. For example, machine learning could be used to develop new diagnostic tools, such as mobile apps that can allow people to take pictures of their skin and receive a diagnosis. Machine learning could also be used to develop new treatments for skin diseases.

1. **Problem statement**

There is a need for more efficient and accurate methods for diagnosing skin diseases.

1. **Project objectives**

Develop a machine learning model that can accurately diagnose skin diseases from images.

1. **Data warehouse details**

The data we used came from a free open-sourced website called Kaggle. The URL to the website is <https://www.kaggle.com/datasets/subirbiswas19/skin-disease-dataset>. It has a public domain license, which means that it free to use by anyone, and there is no restriction (Research guides @ fordham, 2023).

The dataset was collected using internet search. The images of the different skin diseases were then put in labeled folders, with the cause of the skin disease and the name e.g., bacterial infections-cellulites. There is a total of eight skin diseases. The different skin diseases were also then put into two folders, one test\_set folder and train\_set folder. The train\_set folder has the highest number of images with minimun 100 images. The test\_set folder has a minimun of 30 pictures. There is a total of 1159 images of the different skin diseases in the dataset.

Since the dataset already was divided into train\_set folders and test\_set folders we downloaded the folders as they were. But we thought that there might not be enough images to train the models on. The more images the models have to train on the better the accuracy. Which is why we decided to augment the data so there would be more images for the model to train on. Furthermore, the augmented data may be more beneficial for the machine learning model because it can prevent overfitting and it can make the model work well on unseen data.

**5.1 Ethical considerations**

The ethical use of the data is important. The data we are using are pictures of skin diseases. Skin diseases can also appear in faces, which can cause a privacy issue. The images used to train and test the machine learning model we have has no copyright to it, but still there is important to think about the ethics and privacy. We have looked through the images of the data and the data looks anonymous because there are no full pictures of faces. For example, in the folders with skin disease of impetigo, there are pictures of faces, but not of the eyes, which makes it impossible to see who that person is.

Also, let’s say that we want new images to train and test the machine learning model on, then it is important to get permission from the individual so that we can use it in our machine learning model. Furthermore, if this model was made into an app, and people could take pictures of the affected area to classify the skin disease, then it is important that the data is not stored or used without permission.

There are also ethical considerations when there comes to bias in the dataset. We have looked through the data and there seems to be mostly images of white people, this can make the model biased against people of color. This can cause an issue with the accuracy of the model. For example, ringworm, which is one of the skin diseases we are training the model on can appear different on darker skin. In light skin the ring is usually red or pink, but it is usually brown or gray on darker skin (Higuera, 2023).

Overall, while the data is open access there is still important to keep the ethical considerations in mind. Privacy and accuracy are crucial when dealing with healthcare-related datasets.

1. **Interesting insights**
2. **Data pre-processing**
3. **Measuring performance**
4. **Algorithms applied**

**9.1 Algorithm 1**

**9.2 Algorithm 2**

**9.3 KNN**

The K-nearest neighbor algorithm is a machine learning technique used for both classification and regression tasks. The KNN algorithm uses lazy learning to either perform numeric prediction or classification. We have both eager learning and lazy learning. Lazy learning can be referred to as instance-based learning since the learning is based on stored instances of data. While this method is computationally demanding during classification, it can effectively model complex decision spaces that other approaches may not.

The KNN algorithm operates under the assumption that each data point is represented by a tuple in an n-dimensional pattern space, where n corresponds to the number of features in the tuple. Upon encountering a new tuple for classification, the algorithm scans the pattern space to identify the k nearest neighbors to the tuple. For classification tasks, the tuple is assigned the class that is most prevalent among these neighbours. In a numeric prediction task, the output is determined by averaging the numeric values of these neighbor. The value of k, representing the number of neighbor contributing to the similarity assessment, can be optimized through a process of incremental experimentation. The determination of the closest neighbors relies on distance calculations, such as Euclidean distance. Euclidean distance involves calculating the squared difference between corresponding features in two tuples, squaring the result, and adding it to the squared difference of the same calculation for all remaining features. The square root of the sum of these calculations represents the Euclidean distance.

Here is an image that show how KNN works.

Et bilde som inneholder tekst, skjermbilde, diagram, design

Automatisk generert beskrivelse

Image 1:

("K-Nearest Neighbor Algorithm for Machine Learning," n.d .,<https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>)

**Advantages of K-nearest neighbors**

KNN is easy to use because it only needs to find the difference between two points based on their features. This difference can be easily found using formulas like the Euclidean distance or the Manhattan distance (Soni,2020). Furthermore, KNN doesn't need a training phase because the data itself is like a model that's used to make predictions. This makes KNN very fast when it comes to testing new data (Soni,2020). Additionally, since the model doesn't need to be trained, new data can be added anytime. This is because new data won't change the model (Soni,2020).

**Disadvantages of K-nearest neighbors**

While KNN is easy to implement, the efficiency and speed decrease as the dataset size increases (Genesis, 2018). Furthermore, a challenge with KNN is choosing the best number of neighbors to consider when classifying new data point and KNN is easily affected by outliers because it relies on distance when selecting neighbors (Genesis,2018). KNN can't make a general model that's easy to use on new data. It doesn't create rules or charts that can be used to quickly predict new results. Instead, KNN uses all the training data to make its predictions, making it harder to handle large datasets. It also doesn't show how important each piece of information is for making the prediction (MyEducator, n.d).

Overall, the K-nearest neighbors (KNN) algorithm is a simple and powerful machine learning algorithm. It is widely used for classification and regression tasks. It is based on the idea that it is based on the idea that similar data points are likely to have similar labels, and it uses the k nearest neighbors of a new data point to classify or predict its label. The KNN algorithm is well-suited for several applications. It is both simple and efficient and can be a good choice when making a machine learning model. However, it does have some limitations. The algorithm efficiency decreases with larger datasets, and it is sensitive to outliers. Additionally, KNN does not provide any insights into the importance of individual features, which can make it difficult to interpret the results.

**9.4 Comparison**

1. **Discussion**
2. **Conclusion**

**References**

*Recognizing neglected skin diseases: Who publishes Pictorial Training Guide* (2018) *World Health Organization*. Available at: https://www.who.int/news/item/08-06-2018-recognizing-neglected-skin-diseases-who-publishes-pictorial-training-guide (Accessed: 15 November 2023).

*2023 Global Health Care Sector Outlook: Deloitte*, *Deloitte Belgium*. Available at: https://www2.deloitte.com/be/en/pages/life-sciences-and-healthcare/articles/global-health-care-sector-2023.html (Accessed: 16 November 2023).

Džakula, A., Relić, D. and Michelutti , P. (2022) *Health workforce shortage - doing the right things or doing things right?*, *Croatian medical journal*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9086817/ (Accessed: 16 November 2023).

*Healthcare Industry* (2023) *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Healthcare\_industry (Accessed: 17 November 2023).

*Skin diseases: Types of, symptoms, treatment & prevention* (2021) *Cleveland Clinic*. Available at: https://my.clevelandclinic.org/health/diseases/21573-skin-diseases (Accessed: 24 November 2023).

Higuera, V. (2023) *Skin disorders: Pictures, causes, symptoms, and treatment*, *Healthline*. Available at: https://www.healthline.com/health/skin-disorders#causes (Accessed: 24 November 2023).

Federman , D. *et al.* (2004) *A comparison of diagnosis, evaluation, and treatment of patients with Dermatologic Disorders*, *Journal of the American Academy of Dermatology*. Available at: https://www.sciencedirect.com/science/article/pii/0190962295914498 (Accessed: 27 November 2023).

*Shingles* (2023) *Johns Hopkins Medicine*. Available at: https://www.hopkinsmedicine.org/health/conditions-and-diseases/shingles (Accessed: 27 November 2023).

*Cellulitis: All you need to know* (2022) *Centers for Disease Control and Prevention*. Available at: https://www.cdc.gov/groupastrep/diseases-public/Cellulitis.html (Accessed: 28 November 2023).

Dinulos, J.G.H. (2023) *Cutaneous larva migrans - skin disorders*, *Merck Manuals Consumer Version*. Available at: https://www.merckmanuals.com/home/skin-disorders/parasitic-skin-infections/cutaneous-larva-migrans (Accessed: 28 November 2023).

*Research guides @ fordham: Copyright resources: Creative commons & public domain* (2023) *Creative Commons & Public Domain - Copyright Resources - Research Guides @ Fordham at Fordham University Libraries*. Available at: https://fordham.libguides.com/copyright/Licensing (Accessed: 29 November 2023).

“K-Nearest Neighbor Algorithm for Machine Learning." n.d. <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>

Soni, A. (2020) *Advantages and disadvantages of KNN*, *Medium*. Available at: https://medium.com/@anuuz.soni/advantages-and-disadvantages-of-knn-ee06599b9336 (Accessed: 05 December 2023).

Genesis (2018) *Pros and cons of K-nearest neighbors, From The GENESIS*. Available at: https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/ (Accessed: 06 December 2023).

Myeducator (no date) *11.4advantages and disadvantages of KNN, MyEducator*. Available at: https://app.myeducator.com/reader/web/1421a/11/q07a0/ (Accessed: 06 December 2023).