## 

**HACETTEPE UNIVERSITY**

**ENGINEERING FACULTY**

**DEPARTMENT OF COMPUTER ENGINEERING**

**AIN 300**

**INTERNSHIP REPORT**

**Süleyman Yolcu**

**2210765016**

**Performed at**

**TÜBİTAK BİLGEM**

**31/07/2024 - 11/09/2024**

**30 Work Days**

*Report template version: v1. Feb 27, 2019.*

**TABLE OF CONTENTS**

[**1 Introduction**](#_le8rusnetffv) **3**

[**2 Company Information**](#_oj9vt0owaaae) **3**

[2.1 About the company](#_qz3tr64t1du4) 3

[2.2 About your department](#_o3sagc252mym) 3

[2.3 About the hardware and software systems](#_76vp60z8s134) 3

[**3 Work Done**](#_qducuw8epp30) **4**

[**4 Performance and Outcomes**](#_ffl7xu2j0885) **4**

[4.1 Applying Knowledge and Skills Learned at Hacettepe](#_xz50dweoljq9) 4

[4.2 Solving Engineering Problems](#_bpw4ogd4y1rp) 4

[4.3 Teamwork](#_u6lpqhbqxoz3) 4

[4.4 Multi-Disciplinary Work](#_cp0oytm21pkw) 5

[4.5 Professional and Ethical Issues](#_csmm3o9kt8cw) 5

[4.6 Impact of Engineering Solutions](#_g9vfihknoik5) 5

[4.7 Locating Sources and Self-Learning](#_pzho8nnv7pxy) 5

[4.8](#_b6rqjjby5ex1)  [Using New Tools and Technologies](#_ge3qv2cvevyu) 6

[**5 Conclusions**](#_eo3oe5p0knpf) **6**

[**References**](#_l3yfzzp7nwhl) **6**

[**Appendices**](#_b1o40pe12lxb) **7**

# 

# **1 Introduction**

In the introduction section include the following:

* For my internship, I had the privilege of working at TÜBİTAK BİLGEM Bilişim Teknolojileri Enstitüsü (BTE), a prominent research and development center in Turkey focused on information technologies and their applications in various sectors. TÜBİTAK BİLGEM plays a critical role in advancing technological innovation and solutions in fields such as cybersecurity, software development, and data analytics. I chose to pursue my internship at BTE because of its leading role in these areas and its innovative approach to problem-solving through cutting-edge research.
* During my internship, my primary responsibility involved experimenting with different synthetic data generation approaches in healthcare images. This project was essential in addressing one of the key challenges in healthcare: the lack of sufficient and diverse medical data for training machine learning models. By generating synthetic healthcare images, we aimed to improve the quality and quantity of data available for machine learning applications in medical imaging. The work I contributed had direct implications for developing more robust AI models in healthcare, potentially improving diagnosis accuracy and healthcare delivery in the long run.

# **2 Company Information**

## **2.1 About the company**

Provide brief information about the company including the following:

* TÜBİTAK BİLGEM is located in Gebze, Kocaeli
* TÜBİTAK BİLGEM is a part of the TÜBİTAK company. Business areas and keywords
* TÜBİTAK BİLGEM is a national R&D center that produces solutions in the fields of informatics, information security and advanced electronics within the framework of the needs of the civil and military fields and the public and private sectors of the Republic of Turkey.

## **2.2** **About your department**

Describe the department where you did your internship with a focus on the following:

* I worked in TÜBİTAK BİLGEM Cloud Computing and Big Data Research Laboratory
* I worked as an Artificial Intelligence Engineering intern.
* My topic was Synthetic Data Generation techniques.

## **2.3** **About the hardware and software systems**

My department provided me a working laptop and gave accsses to its GPU servers.

**2.4**  **About your supervisor**

Provide the following information for the supervisor:

* Name and surname,
* address,
* telephone number,
* email address, and
* information about the education of your supervisor (including the name of the university and department from which he/she graduated, and the year of graduation).

# **3** **Work Done**

* During my internship at TÜBİTAK BİLGEM Bilişim Teknolojileri Enstitüsü (BTE), I was involved in a project focused on experimenting with different synthetic data generation techniques for healthcare images. The primary goal of this project was to address the shortage of diverse medical datasets, which are crucial for developing accurate machine learning models in the healthcare sector. Specifically, the work aimed to generate synthetic chest X-ray images for the diagnosis of various lung diseases such as pneumonia, tuberculosis, and COVID-19. The synthetic data would help improve the robustness of AI-based diagnostic tools, ensuring better detection accuracy even with limited real-world data.

**Motivation and Importance**

* The project was driven by the need for a larger and more diverse dataset in the medical field, where privacy concerns, cost, and the availability of data are significant challenges. The synthetic data generation would enable researchers to train machine learning models more effectively by providing an expanded set of training images that accurately represent medical conditions without compromising patient privacy. This work contributes to improving AI systems used for diagnosis, making healthcare more accessible and efficient.

**Detailed Description of the Work**

1. Developed Models and Algorithms

* The project involved the development and training of three different machine learning models:
* **Generative Adversarial Networks (GANs)**: GANs consist of two networks—the generator and the discriminator. The generator creates synthetic images, while the discriminator tries to distinguish between real and generated images. The two networks are trained simultaneously, improving the quality of the generated images over time.
* Pseudo-code outline for GAN:
* while training:  
   generator\_output = Generator(noise\_input)  
   real\_vs\_fake = Discriminator(generator\_output, real\_image)  
   calculate\_loss()  
   update\_weights()
* **Variational Autoencoders (VAEs)**: VAEs are used for generating new images by encoding real images into a latent space and then decoding them back into images. They introduce a probabilistic element to ensure diverse synthetic outputs.
* Pseudo-code outline for VAE:
* encode(input\_image)  
  latent\_space = mean\_and\_variance()  
  decode(latent\_space)  
  reconstruct\_image()
* **Denoising Diffusion Probabilistic Models (DDPMs)**: These models generate synthetic images by gradually reducing noise from a noisy image to create a clear image. This novel approach was critical for generating high-quality synthetic medical images.
* Pseudo-code outline for DDPM:
* start\_with\_noise()  
  for each step in reverse\_process:  
   predict\_noise\_removal()  
   update\_image()

2. Data and Preprocessing

* The project used the Chest X-Ray Images (Pneumonia) dataset from Kaggle, which contains 5,216 images of healthy and pneumonia-infected lungs. Preprocessing involved:
* Resizing all images to 64x64 pixels for efficiency.
* Converting to grayscale as X-ray images primarily rely on contrast rather than color.
* Labeling the data into two classes: Normal (healthy lungs) and Pneumonia (infected lungs).

3. Hardware and Software Environment

* Hardware: The models were trained on high-performance GPUs to speed up the training process.
* Software: The project used Python and the following libraries:
* TensorFlow and PyTorch for model development.
* NumPy and Pandas for data manipulation.
* OpenCV for image processing.

4. Design and Testing Methods

* Design Methods: The models were designed based on standard deep learning architectures. The generator and discriminator for the GAN, encoder-decoder for the VAE, and forward-backward processes for the DDPM were carefully adjusted.
* Testing Methods: To evaluate the performance of the generated images, we used:
* Inception Score (IS) to assess the diversity and quality of the synthetic images.
* Frechet Inception Distance (FID) to measure how close the generated images were to the real ones.

5. Project Management and Processes

* Agile Methodology was used throughout the project to iteratively improve the model and adjust parameters based on performance metrics.
* Version Control: All code and data were managed using Git, ensuring that different versions of the models could be tracked and compared.

6. Engineering Standards Followed

* Ethical standards: The work followed data privacy guidelines, ensuring that no real patient information was used improperly.
* Software development standards: Best practices for coding, including documentation and testing, were adhered to throughout the project.

My Contribution

* I was directly responsible for experimenting with the three models—GANs, VAEs, and DDPMs—and adjusting their architectures to produce the best quality synthetic healthcare images. I also handled the preprocessing of the dataset, training the models on GPUs, and testing the results using IS and FID scores.

# **4 Performance and Outcomes**

## **4.1** **Applying Knowledge and Skills Learned at Hacettepe**

During my internship at TÜBİTAK BİLGEM Bilişim Teknolojileri Enstitüsü (BTE), I had the opportunity to apply a variety of concepts and skills that I learned during my studies at Hacettepe University. Here are some specific areas:

* **Machine Learning and Deep Learning Fundamentals**: The core knowledge I gained in machine learning courses helped me understand how to apply models like GANs, VAEs, and DDPMs. Specifically, the concept of neural networks, training processes (forward and backward propagation), and loss functions was critical in developing and improving synthetic data generation models.
* **Data Preprocessing Techniques**: My education in handling data was essential when preparing the medical image dataset. I utilized methods like image resizing, normalization, and converting images to grayscale, which are techniques I learned in my courses on data science and image processing.
* **Programming and Software Tools**: Throughout my studies, I gained proficiency in Python and its associated libraries, such as TensorFlow, PyTorch, and OpenCV. These tools were heavily utilized during my internship for both model development and data manipulation. Additionally, the software engineering practices I learned, such as using Git for version control, played a key role in managing the project's progress.
* **Mathematical Foundations**: The knowledge of linear algebra, probability, and optimization techniques, which I studied in mathematics courses, was crucial for understanding and tweaking the architectures of the deep learning models, particularly in handling latent space representations and calculating probabilistic distributions in VAEs.

## **4.2** **Solving Engineering Problems**

During the internship, I encountered and solved several engineering problems, particularly related to the development and optimization of machine learning models and data handling in healthcare applications. Here are some examples:

* **Model Convergence and Stability**: One of the main problem was the instability in the training process of the GAN model, which could result in poor-quality synthetic images. By carefully adjusting hyperparameters, such as learning rate and batch size, and fine-tuning the architecture of the generator and discriminator, I was able to stabilize the training process. This led to improved convergence, producing more realistic synthetic images.
* **Balancing Computational Efficiency**: Training deep learning models on large datasets can be computationally expensive and time-consuming. To address this, I applied techniques such as mini-batch gradient descent and used GPUs to speed up the training process. By optimizing the code and utilizing hardware effectively, I was able to reduce the time required for model training without compromising performance.
* **Ensuring Quality of Synthetic Data**: A major challenge was to ensure that the synthetic images generated by the models were both realistic and diverse enough to be useful for training AI models. I solved this problem by implementing evaluation metrics such as Inception Score (IS) and Frechet Inception Distance (FID). These metrics allowed us to measure the quality and diversity of the synthetic images and iteratively improve the models to achieve the desired outcomes.

## **4.3** **Teamwork**

Explain in detail the teamwork you were involved in during the internship, including (for each team you participated in)

* the team role of each team member,
* the training in their background and
* current work area, and
* some information about the team dynamics as you worked together.

You should clearly explain how you related to the others on the team. If you were not involved in a formal team, the definition of the term could be interpreted loosely to mean working together with others on a shared task.

**4.4**  **Multi-Disciplinary Work**

Explain how you worked with team-mates from other disciplines other than computer engineering. Give the details of the multi-disciplinary work that you have carried out as a team. If you were not a part of a multi-disciplinary work or team during your internship, indicate and explain this in this section clearly.

## **4.5** **Professional and Ethical Issues**

I did not encounter any significant professional or ethical issues. The organization follows strict ethical guidelines, particularly when it comes to data privacy, which is a critical concern in healthcare. Since we were working with synthetic data, we ensured that no real patient data was compromised, which helped mitigate ethical risks related to privacy and confidentiality. The company's commitment to these ethical standards was evident throughout the project, and the entire team adhered to industry best practices regarding the use of sensitive data.

## **4.6** **Impact of Engineering Solutions**

One of the key realizations I had during my internship was how engineering solutions can have a profound economic, societal, and global impact, particularly in the healthcare sector. The project I worked on—generating synthetic healthcare images—has the potential to significantly lower the cost of healthcare diagnostics by providing more robust AI models without the need for vast amounts of real patient data. These models could reduce the burden on healthcare providers by improving diagnostic accuracy, which could lead to quicker treatments and better patient outcomes. Moreover, by reducing the dependency on real medical data, the project addressed privacy concerns, thus contributing to the ethical use of AI in healthcare.

## **4.7** **Locating Sources and Self-Learning**

Throughout the internship, I engaged in significant self-learning to deepen my understanding of the tools and methodologies used in the project. I relied heavily on online platforms such as YouTube and Coursera, where I found tutorials and courses on advanced topics related to synthetic data generation, deep learning, and model optimization. These resources were crucial in helping me better understand complex topics like GANs and the underlying mathematics of machine learning.

For example, I used YouTube to find practical coding examples for implementing the Generative Adversarial Networks (GANs) and Coursera for deeper theoretical knowledge of machine learning frameworks and best practices. These resources allowed me to quickly overcome challenges I faced during the internship and contributed to my ability to solve real-world problems effectively.

## **4.8** **Using New Tools and Technologies**

During my internship, I encountered GPU servers for the first time, which were essential for the efficient training of deep learning models. Before this internship, I had only worked with CPUs, and GPU servers provided a substantial increase in computational power, allowing for faster training of large-scale machine learning models.

I learned how to set up, configure, and optimize the use of GPUs for model training through online tutorials and guidance from my supervisors. By the end of the internship, I became proficient in utilizing GPU resources to optimize the training time of deep learning models, a skill that significantly enhanced my technical abilities and will be beneficial for future projects involving large datasets and computationally heavy tasks.

**5 Conclusions**

During my internship at TÜBİTAK BİLGEM Bilişim Teknolojileri Enstitüsü (BTE), I had the opportunity to contribute to a project focused on experimenting with different synthetic data generation approaches for healthcare images. The primary goal was to generate synthetic chest X-ray images to improve the availability and diversity of medical data, which is crucial for training AI models in the healthcare field. My work involved developing and fine-tuning models like Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Denoising Diffusion Probabilistic Models (DDPMs), all of which were instrumental in producing realistic and high-quality synthetic images.

Throughout this experience, I applied the knowledge and skills I gained from my studies at Hacettepe University, particularly in areas such as machine learning, and data preprocessing techniques. Overall, my contribution to the project involved improving the training process and the quality of synthetic data generation models, which could significantly impact healthcare diagnostics by providing robust AI models that require less real-world data

# **References**

Give proper citations to all of the resources you have used during your internship. When giving references, be sure to adhere to ACM or IEEE reference format. The reference examples should adapted from ACM and IEEE reference style suggestions that can be found at ACM and IEEE websites [1,2].

[1] “IEEE Reference Guide”. http://ieeeauthorcenter.ieee.org/wp-content/uploads/IEEE-Reference-Guide.pdf. [Accessed; Feb 27, 2019].

[2] “ACM Citation Style and Reference Formats”. https://www.acm.org/publications/authors/reference-formatting. [Accessed: Feb 27, 2019].

# **Appendices**

If necessary, provide any detailed information such as long code examples, extensive company information, images or results charts.