Text Summarization using Deep Learning

Aditya Anekar

Jaimik Patel

Srishti Shankar

Yug Dave

SDS-384 Scientific Machine Learning

Prof. Arya Farahi

Spring 2024

The University of Texas at Austin

Index

Chapter 1: Introduction

In the age of information overload, where vast quantities of data, articles, scientific papers, and online content are generated daily, it becomes increasingly challenging to consume and comprehend all available information efficiently. Text summarization emerges as a critical solution to this problem, serving to condense the extensive texts into concise summaries without sacrificing their core messages. This technique not only facilitates a quicker understanding of the main points but also significantly saves time, making information consumption more manageable.

Text summarization can be approached through various methods like extractive or abstractive. We will be using abstractive deep learning techniques. Unlike extractive summarization, which simply picks and combines chunks of the original text, abstractive summarization leverages deep learning to understand the context and generate entirely new sentences by paraphrasing or rephrasing the original content. This approach enables the creation of more natural and less redundant summaries, closely mimicking human-like comprehension and synthesis.

The motivation for developing a text summarization model stems from the sheer information overload on the internet due to the huge number of articles, scientific papers, reports & content present that professionals across fields must navigate. For instance, researchers and academics often face the daunting task of keeping up with the ever-expanding corpus of scientific literature. Similarly, business professionals and policymakers require swift assimilation of lengthy reports and documents to make informed decisions. In all these scenarios, abstractive text summarization can dramatically reduce reading time, allowing for the distillation of essential information and making the vast seas of data navigable.

Hence, the development of abstractive text summarization models is not merely a technical challenge but a necessity in our current digital landscape. By employing deep learning algorithms, we will build a system that not only summarizes texts efficiently but also enhances the accessibility and usability of information, paving the way for more informed and timely decision-making in an increasingly complex world.

Chapter 2: DataSet

The dataset that we will be using for this project is the PubMed dataset, accessible via the Hugging Face datasets library[11]. The dataset has three columns namely article, abstract, and section. Each entry in the dataset includes the full text of the document under the 'article' field. The 'abstract' field contains the document's abstract, serving as a summary ideal for training summarization models. Additionally, the 'section_names' field lists the titles of various sections within the document, offering insights into the document's structure and key focus areas. The dataset contains a total of 133,000 entries which we divided into 120,000 training samples, 6,600 validation samples, and 6,300 test samples, for robust model training and evaluation.

	Unnamed: 0	article	abstract	section_names
0	73434	to review the presentation and histological di...	objective : to review the presentation and hi...	Objective:\nMaterials and Methods:\nResults:\n...
1	7459	ethylcellulose , a nonbiodegradable and biocom...	\n objective . \n the purpose of the recent s...	1. Introduction\n2. Materials and Methods\n3. ...
2	136	acute generalized exanthematous pustulosis (a...	acute generalized exanthematous pustulosis (...	Introduction\nCase Report\nDiscussion
3	76845	\n physical restraint is a coercive interventi...	\n background : considering the negative cons...	Introduction\nMethods\nResults\nDiscussion\nCo...
4	80361	disease registries are considered reliable sou...	the main aim of this study is to determine th...	1. Introduction\n2. Materials and Methods\n3. ...

Figure 1: Dataset snippet

We implemented a systematic approach to ready the PubMed dataset for training our T5-Small and BART-Base models. The process involved several key steps:

- Lowercasing: All text within the corpus was converted to lowercase to ensure consistency and uniformity, thereby preventing word duplication due to capitalization discrepancies.
- Length-based Filtering: To exclude overly lengthy or extremely short texts that might not contribute significantly to the training, we employed the interquartile range (IQR) method. Texts falling outside a certain length range were eliminated from the dataset, with thresholds set at 1.5 times the IQR above the third quartile and 1.5 times the IQR below the first quartile.
- Preservation of Special Characters: Special characters, symbols, and punctuation marks were retained within the text. Given the scientific nature of the PubMed dataset, these characters often convey crucial information such as chemical formulas, gene names, and mathematical symbols. Hence, it was imperative to maintain their presence to uphold the scientific integrity of the content.

Chapter 3a: Exploratory Analysis

1. Length of the articles vs abstracts in the dataset

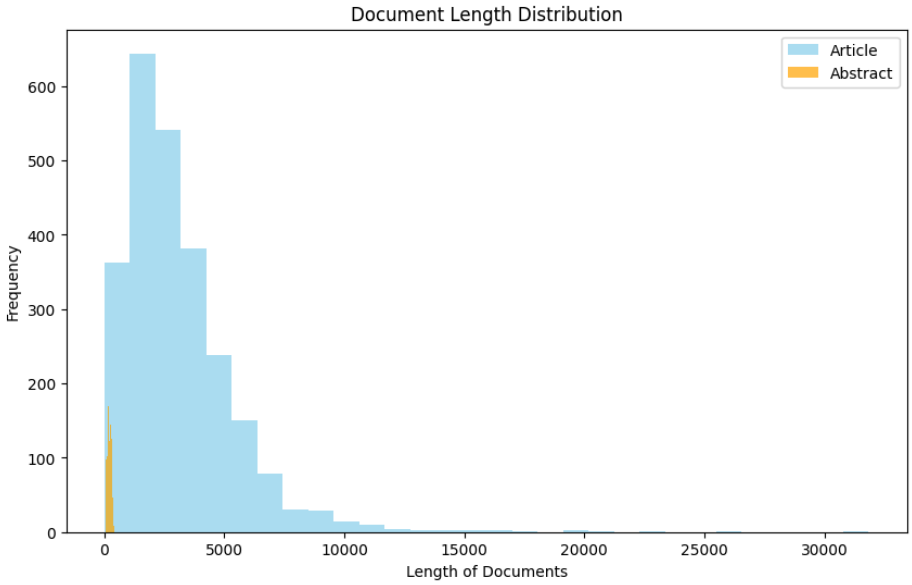


Figure 2: Document Length Distribution

For the first exploratory analysis, we wanted to compare the length of the article and the abstract after summarization. Therefore, in Figure 2, we can see that most of the articles are longer whereas the abstracts are considerably smaller, with maximum length hardly even reaching 500 words.

2. Similarity Analysis Between Abstracts and Articles

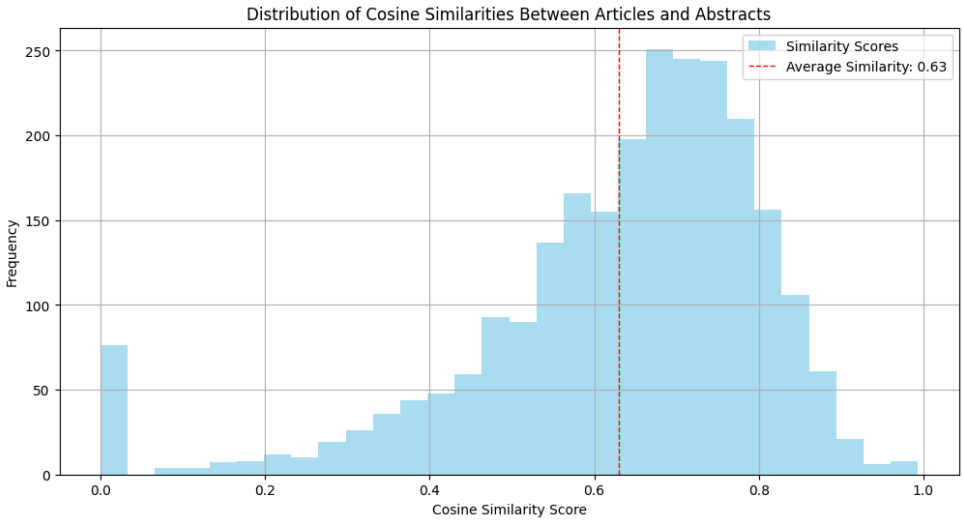


Figure 3: Cosine Similarity between Articles and Abstract

Next, we wanted to check the similarity between the articles and the abstract in the dataset. Therefore, we convert text to numerical vectors using TF-IDF and then calculate the cosine similarity between the abstract and article text of each record. Here the cosine similarity score = 1, would signify perfect similarity between the article and abstract. We can see in Figure 3, that the average similarity score is 0.63 between the article and abstract with more concentration between 0.6 and 1.0.

3. The sentiment analysis between the article and the abstract

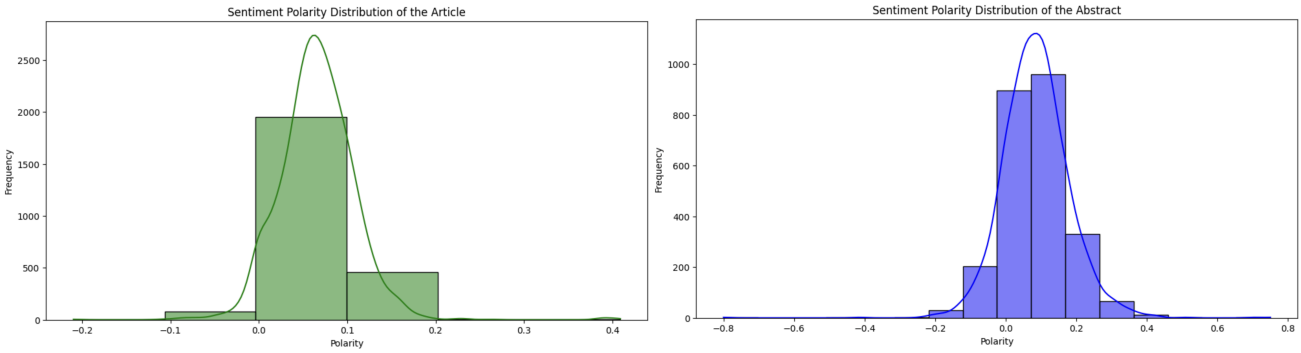


Figure 4: Sentiment Polarity Distribution

In Figure 4, we wanted to analyze the sentiments between the article and the abstract, therefore we calculated the sentiment polarity of both of them. We can see the sentiment polarity score between the article and the abstract both are somewhat peaking between 0.0 and 0,1. This signifies that they both convey the same message and are similar in terms of the sentiment of the message.

4. Most common words in the Abstract.

Figure 5: Common words in Abstract

We conducted another analysis to see the most common words present in the abstract and from the above figure 5, it was clear that “patients” we used the most, followed by “study”, “group” and “treatment”.

Chapter 3b: Hypothesis Testing

1. The lexical diversity in the abstracts is significantly lower than in the full articles due to the concise and focused nature of abstracts.

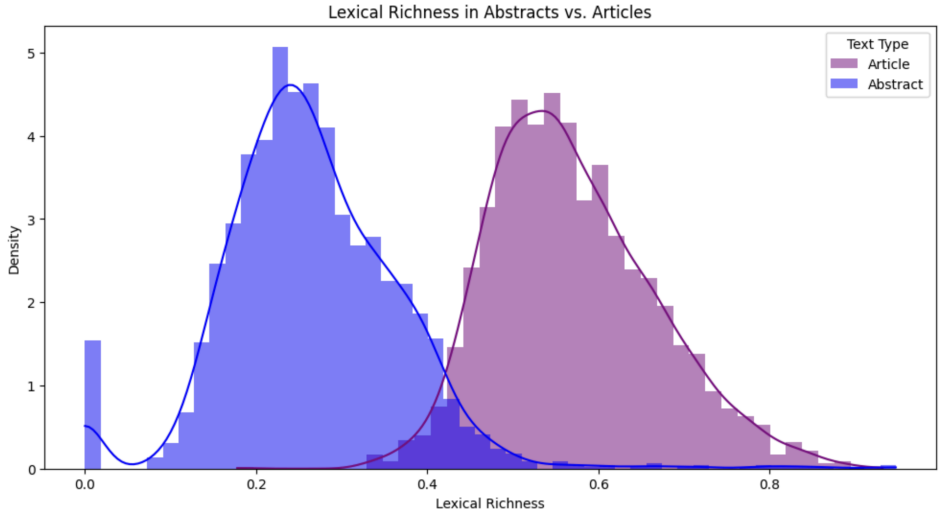


Figure 6: Lexical Richness plot

Abstracts are short summaries of articles, giving a quick overview of the main points. They are brief, so they usually do not use as many different words as the full articles.

We measured *Lexical Richness* by forming sets of unique words and comparing them to the overall word count of the abstract or the article. In Figure 6, we can see visually that abstracts do not use as many different words as the full articles. This suggests that abstracts are more focused and use fewer unique words than full articles.

2. There is a positive correlation between the length of the article and the length of its abstract.

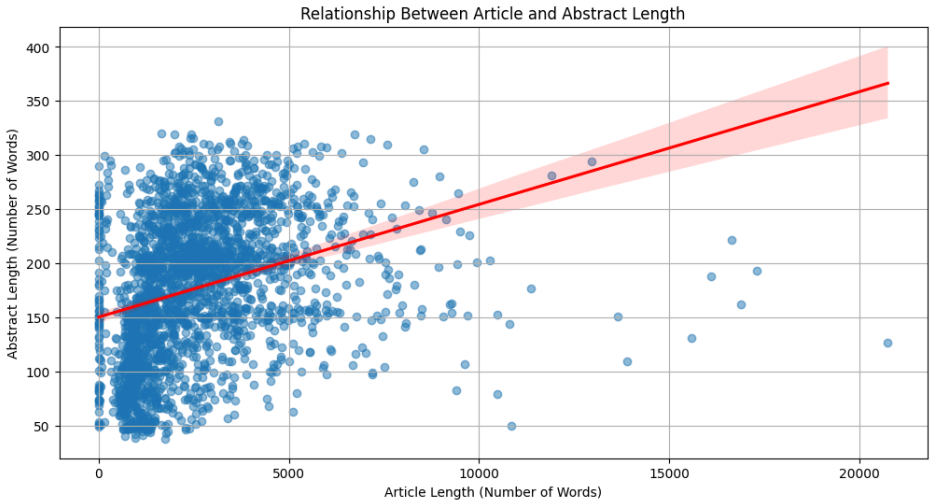


Figure 7: Article length vs Abstract length

The size of an article seems to affect how long its abstract is. In Figure 7, we can see a scatter plot with a line that shows how they are related. Longer articles usually have longer abstracts, meaning they need more words to summarize more detailed discussions. The line's slope in the plot tells us how strong this relationship is: as articles get longer, their abstracts also tend to get longer by about the same proportion.

3. The detail level of abstracts (measured by the variety of unique words) have a greater similarity score with the article (detailed abstracts are more closely aligned with the full content of the article).

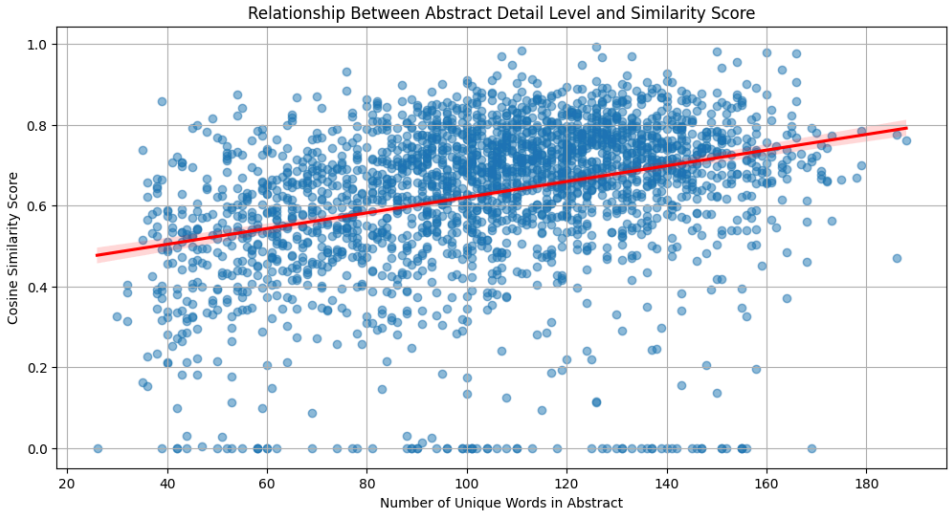


Figure 8: Unique words in Abstract vs Similarity score

The complexity of an abstract, which we can figure out by looking at how many different words it uses, might show how well it matches the article it summarizes.

We measure the similarity by using the *Cosine Similarity Score* and compare it with *unique* abstract words. In Figure 8, we see that abstracts with more unique words usually match their articles better. This means that more detailed abstracts, with a wider variety of words, are likely to capture the main ideas of the article more accurately. They mirror the article's overall subject matter more closely.

4. Articles with more sections have a lower similarity score between their abstracts and articles.

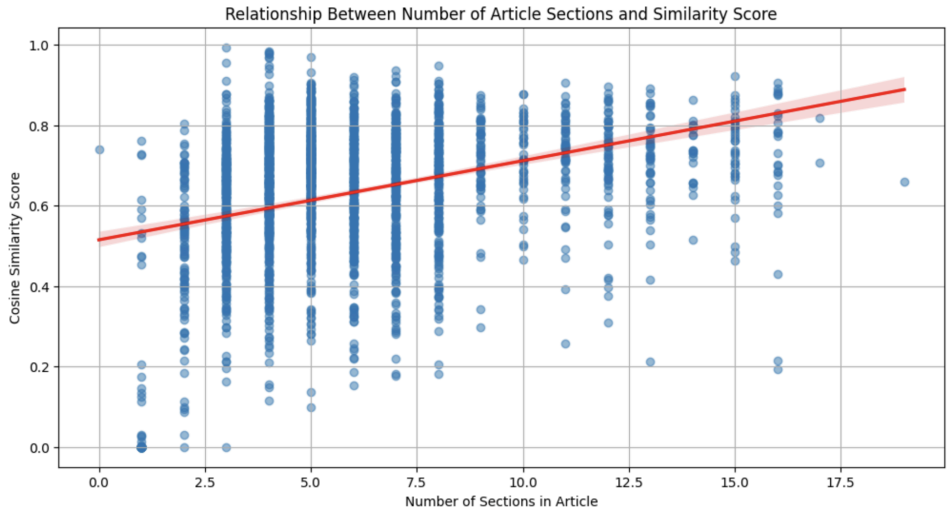


Figure 9: Number of sections in article vs Similarity score

Figure 9 shows something disproving our hypothesis. Articles with more sections tend to have abstracts that are more similar to the full article, which is surprising. One way to understand this is that when articles have many sections, they are often well-organized, making it easier to see the main themes and topics. This organized structure might help when writing the abstract, making it easier to capture and reflect on the main points of each section. So, abstracts can summarize the key points from each section more easily, which keeps them closely aligned with the overall content of the article.

Chapter 4a: Modeling

Model Architecture

T5 Model

The *T5 (Text-To-Text Transfer Transformer)* model is a state-of-the-art neural network architecture for a wide range of natural language processing tasks, including text summarization. T5 models are based on the Transformer architecture, which has shown remarkable success in various NLP applications due to its attention mechanism and self-attention mechanism. Key details of the model are as follows:

- Transformer Architecture: Like other models in the Transformer family, T5-Small consists of multiple layers of self-attention mechanisms and feed-forward neural networks. Each layer has two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network.
- Embeddings: T5-Small utilizes token embeddings to convert input tokens into fixed-dimensional vectors. These embeddings capture semantic information about the input tokens and serve as the initial representation of the input sequence.
- Encoder-Decoder Structure: T5-Small follows an encoder-decoder architecture, where the input text is first processed by the encoder to generate contextualized representations, and then the decoder generates the output sequence based on these representations.
- Text-To-Text Format: One of the key features of the T5 model is its text-to-text approach, where all tasks, including summarization, are framed as text generation tasks which allows T5 models to handle various NLP tasks with a single architecture.
- Task-Agnostic Pretraining: T5-Small is pre-trained on a large corpus of text data using a task-agnostic objective which allows the model to learn general linguistic patterns and representations that can be fine-tuned for specific downstream tasks.
- Parameter Size: T5-Small has a relatively small number of parameters compared to larger variants, making it more suitable for environments with limited computational resources. Despite its smaller size, the T5-Small retains much of the expressiveness and capabilities of larger models.

BART Model

BART (Bidirectional and Auto-Regressive Transformers) is a neural network architecture introduced by Facebook AI Research (FAIR) that combines the bidirectional encoder architecture with an auto-regressive decoder. It is specifically designed for sequence-to-sequence tasks, including text summarization, generation, and translation. BART is built upon the Transformer architecture and

incorporates several innovative techniques to enhance its performance and efficiency. Key model architecture details are as follows:

- Bidirectional Encoder: Similar to other Transformer-based models, BART consists of an encoder-decoder architecture. The encoder processes the input sequence bidirectionally, capturing contextual information from both the left and right contexts of each token.
- Auto-Regressive Decoder: BART employs an auto-regressive decoder to generate the output sequence token by token.
- Masked Self-Attention Mechanism: BART uses masked self-attention mechanisms in both the encoder and decoder layers. In the decoder, self-attention ensures that each token is generated based only on the previously generated tokens.
- Input and Output Token Embeddings: BART utilizes token embeddings to convert input and output tokens into fixed-dimensional vectors.
- Pre-Training Objective: BART is pre-trained using a denoising autoencoding objective, where corrupted input sequences are reconstructed to their original form.
- Parameter Size: The BART Base model is a medium-sized variant of the BART family, with a moderate number of parameters. It strikes a balance between model size and performance, making it suitable for a wide range of applications while still being computationally efficient.

Preprocessing

Data Cleaning

In the preprocessing phase of our project, we adopted a systematic approach to prepare the PubMed dataset for training our T5-Small and BART-Base models. The following steps were undertaken:

- Lowercasing: The entire text corpus was converted to lowercase to ensure uniformity and consistency in the text data. This step helps prevent the duplication of words due to differences in capitalization.
- Filtering Based on Length: To filter out excessively long or short texts that might not contribute meaningfully to the training process, we utilized the interquartile range (IQR) method. Texts falling outside a certain range of lengths were removed from the dataset. (1.5*IQR above third quartile and 1.5*IQR below first quartile)
- Retaining Special Characters: Special characters, symbols, and punctuation marks were retained in the text. Given the scientific nature of the PubMed dataset, special characters often carry important information, such as chemical formulas, gene names, and mathematical symbols. Thus, preserving these characters was essential to maintain the integrity of the scientific content.

Tokenization

After cleaning the dataset, we proceeded with the tokenization process using the respective tokenizers for T5-Small and BART-Base models. Tokenization involves breaking down the text into smaller units, typically words or subwords, which serve as input tokens for the model. The following steps were involved:

- T5-Small Tokenization: For the T5-Small model, we utilized the tokenization scheme specifically designed for T5 architectures. This tokenizer breaks the input text into subword units and assigns a unique token ID to each subword. Additionally, special tokens, such as <pad>, <bos>, <eos>, and <unk>, were added to facilitate model training and decoding.
- BART-Base Tokenization: Similarly, for the BART-Base model, we employed the tokenizer tailored for BART architectures. This tokenizer also employs a subword tokenization approach, capturing the morphological and syntactic structure of the input text. Special tokens specific to BART, such as <s> (start of sequence) and <pad>, were included to delineate the beginning and padding of sequences.

Fine-Tuning Process

Training Setup

The fine-tuning process of the T5-Small and BART-Base models was conducted using the following setup:

- Training Environment: The training was conducted on the lightning.ai platform, leveraging its infrastructure for distributed training capabilities. The environment provided 8 virtual CPUs (vCPUs) and 1 Nvidia L4 GPU, enabling efficient parallel processing of training data.
- Training Arguments: The most important training arguments are explained below.
    - **Number of Training Epochs:** The number of times the entire training dataset is passed through the model during training.
    - **Per-Device Batch Size:** The batch size for each device (GPU) during training and evaluation.
    - **Warmup Steps:** The number of steps used for learning rate warm-up during training.
    - **Weight Decay:** The coefficient for L2 regularization to prevent overfitting.
    - **Early Stopping:** If validation loss during training fails to decrease for 3 consecutive steps, then training stops early.

Hyperparameters

```
# Define the training arguments
training_args = TrainingArguments(
    output_dir='./results',
    num_train_epochs=3,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=4,
    warmup_steps=500,
    weight_decay=0.01,
    learning_rate=5e-5,
    logging_steps=100,
    save_steps=100,
    eval_steps = 100,
    save_total_limit = 3,
    load_best_model_at_end=True,
    evaluation_strategy = IntervalStrategy.STEPS,
    metric_for_best_model="eval_loss",
    greater_is_better=False
)

# Initialize the Trainer
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_tokenized_datasets,
    eval_dataset=valid_tokenized_datasets,
    callbacks = [EarlyStoppingCallback(early_stopping_patience=3)]
)
```

Figure#: Screenshot of Model Arguments (Both models were trained using the same arguments)

Evaluation Metrics

A comprehensive set of evaluation metrics were employed to measure the performance of the model:

- **ROUGE Scores:** These metrics measure lexical and structural similarity between the generated and reference summaries, providing insights into content overlap and linguistic fluency.
- **BERT Score:** Utilizing contextual embeddings from BERT, this metric evaluates the semantic similarity between the generated and reference summaries, offering a nuanced assessment beyond simple token overlap.
- **BLEU Score:** Assessing *n-gram* precision, this metric evaluates the adequacy and fluency of the generated summaries by comparing them to reference summaries.
- **METEOR Score:** This metric considers linguistic and semantic features such as synonymy and paraphrasing, providing a comprehensive evaluation of summary quality beyond token overlap.
- **Cosine Similarity:** Quantifying semantic similarity in vector space, this metric offers insights into the coherence and relevance of the generated summaries by measuring the angle between their vectors.

Training Performance

T5-Small Model

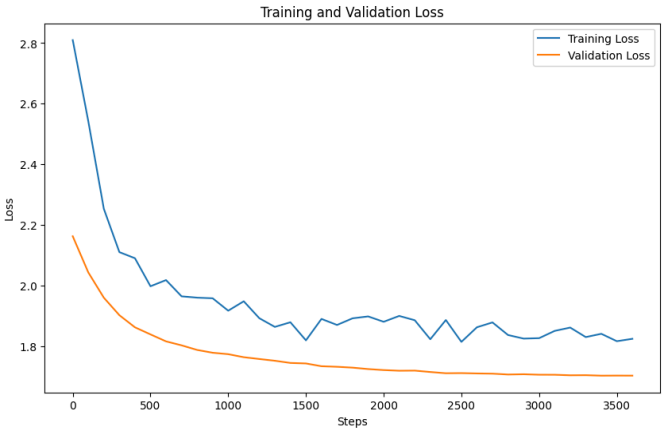
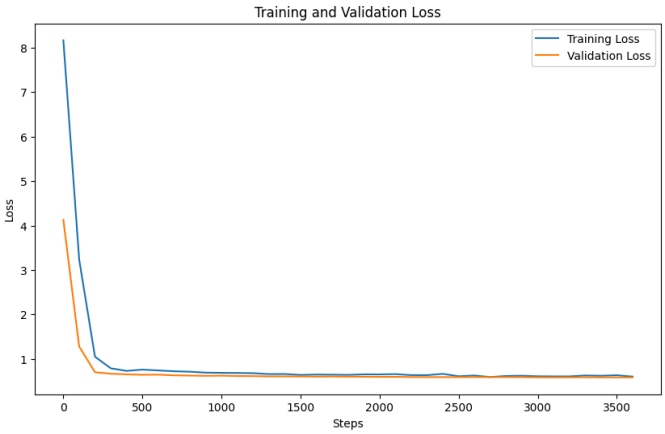


Figure #:

The model seems to be performing well, as indicated by the decreasing loss values. However, the slight fluctuations in the validation loss suggest that there might be some variance in the model's performance on different parts of the dataset. This graph does not show an overfitting trend, which is positive. In this graph, since both losses are decreasing, it suggests that the model has not yet reached its full potential and could benefit from further training.

BART-Base Model



Figure#:

The fact that the Training Loss is consistently below the Validation Loss is a good sign. It indicates that the model is not overfitting to the training data and is likely to generalize well to new, unseen data. The plateauing of the loss values suggests that the model may have reached its learning capacity with the current data and architecture. To achieve further improvements, one might consider increasing the complexity of the model or adding more diverse training data.

Chapter 4b: Validation

Data Splitting Strategy

The dataset was split into three subsets: training, validation, and test sets. The training set was used to train the model, the validation set was used to tune hyperparameters and monitor model performance during training, and the test set was used to evaluate the final performance of the trained model. The data was split according to a ratio of **90%** for *training*, **5%** for *validation*, and **5%** for *testing*. This ratio was chosen to allocate the majority of the data for training while ensuring sufficient data for validation and testing to obtain reliable performance estimates. The relatively small size of the validation and test sets helps ensure they are representative dataset samples while still providing robust evaluation.

Early Stopping

Validation loss was monitored during model training to gauge performance on unseen data. Early stopping was used to halt training if validation loss ceased to decrease for 3 consecutive logging steps. Early stopping prevents overfitting by halting training before the model memorizes the training data excessively. Halt training when validation loss stagnates saves computational resources and time. Early stopping ensures the model selection with the best performance on unseen data, enhancing model quality.
We utilized the Hugging Face Transformers library's *EarlyStoppingCallback* to monitor validation loss. The training was halted if validation loss did not decrease for three consecutive logging steps.

Overfitting Analysis

To prevent overfitting, a weight decay of 0.01 was applied as a form of regularization during training. Additionally, other regularization techniques provided by the Hugging Face Transformers library, such as dropout and layer normalization, were utilized implicitly during model training.
Despite the absence of explicit overfitting measures such as cross-validation, precautions were taken to mitigate overfitting through regularization techniques. The impact of weight decay and other regularization techniques on model generalization and performance was analyzed through validation metrics and convergence analysis.

Chapter 5: Results

Quantitative Evaluation

Metrics

On fine-tuning the model on the subset of the original dataset, the following scores were obtained.

Table#: Scores

Score(Model)	Precision	Recall	F1	Score
Rouge-1 (T5)	0.41372	0.24552	0.29115	
Rouge-1 (BART)	0.46875	0.30029	0.34293	
BLEU (T5)				0.09512
BLEU (BART)				0.09538
BERT (T5)	0.83440	0.80646	0.81987	
BERT (BART)	0.85203	0.83545	0.84339	
METEOR (T5)				0.14111
METEOR (BART)				0.22519
Cosine (T5)				0.54
Cosine (BART)				0.65

Qualitative Evaluation

Original Long Text

Refer to the appendix for the full text of the example article.

Reference Summary

“background . bronchopulmonary dysplasia (bpd) is the most common chronic lung disease of premature birth , characterized by impaired alveolar development and inflammation . pathomechanisms contributing to bpd are poorly understood . however , it is assumed that genetic factors predispose to bpd and other pulmonary diseases of preterm neonates , such as neonatal respiratory distress syndrome (rds) . for association studies , genes upregulated during alveolarization are major candidates for genetic analysis , for example , matrix metalloproteinases (mmps) and fibroblast growth factors (fgfs) and their receptors (fgfr) . objective . determining genetic risk variants in a caucasian population of premature neonates with bpd and rds . methods . we genotyped 27 polymorphisms within 14 candidate genes via

restriction fragment length polymorphism (rflp) : mmp-1 , -2 , -9 , and -12 , -16 , fgf receptors 2 and 4 , fgf-2 , -3 , -4 , -7 , and -18 , signal - regulatory protein (sirpa) and thyroid transcription factor-1 (ttf-1) . results . five single nucleotide polymorphisms (snps) in mmp-9 , mmp-12 , fgfr-4 , fgf-3 , and fgf-7 are associated (p < 0.05) with rds , defined as surfactant application within the first 24 hours after birth . one of them , in fgfr-4 (rs1966265) , is associated with both rds (p = 0.003) and bpd (p = 0.023) . conclusion . rs1966265 in fgf receptor 4 is a possible genetic key variant in alveolar diseases of preterm newborns ."

Representative Samples Generated by the Models

T5

"bronchopulmonary dysplasia (bpd) remains a major complication of premature birth. despite these advances, 45 years after its first description, the old bpd is mainly affected modestly premature newborns suffering from respiratory distress and therefore mechanically ventilated with high levels of supplemental oxygen. despite this advances, 45 years after its first description, prenatal maternal use of glucocorticoids, improved nutrition, and ventilator strategies the clinical course and pathology of bronchopulmonary dysplasia isoforms are important determinants for alveolarization."

BART

"background and objectives : bronchopulmonary dysplasia (bpd) remains a major complication of premature birth, causing ongoing morbidity and mortality. the purpose of this study was to identify genetic risk factors in a caucasian population of premature newborn with bpd and rds.materials and methods : we genotyped 27 polymorphisms within fourteen candidate genes for bpd : mmp-1, -2, -9, -12, and -16, fibroblast growth factor (fgf), -3, -4, -7, -18, and matrix metalloproteinase (mmp -9), which are important determinants for alveolarization in developing lungs and to protect from bpd.results:a total of 27 genotypes of bpd were identified, with a mean age ranging from 2 to 28 weeks of gestation (p < 0.001). of the genotypes, there was a significant association between mmp2 and -9 (p = 0.002) and bpd. in addition, there was no association between bpd or rds."

Insights

The summary generated by the T5 model contains some relevant information about BPD, mentioning its status as a complication of premature birth and its impact on alveolarization. However, the summary lacks coherence and clarity. It repeats the phrase "despite these advances" twice, which seems redundant and does not contribute to the overall coherence of the summary. Additionally, some phrases are disjointed and do not flow smoothly, impacting the readability of the summary. Overall, while the T5 model captures

some aspects of the reference text, its summary lacks coherence and does not effectively convey the main points.

In contrast, the summary generated by the BART model closely mirrors the structure and content of the reference text. It provides a comprehensive overview of BPD, its characteristics, and the genetic factors associated with it. The summary is well-structured and coherent, maintaining the logical flow of information present in the reference text. It effectively captures key concepts and findings, such as the association between genetic variants and BPD, without unnecessary repetition or ambiguity. Overall, the BART model produces a high-quality and coherent summary that closely aligns with the content and structure of the reference text.

Chapter 6: Discussion

Interpretation of Results

- **BART's Superior Alignment with Reference Summaries:** BART generally achieves higher scores across most metrics compared to T5. This suggests that summaries generated by BART are more aligned with the reference summaries in terms of both the presence of specific words and the overall meaning.
- **Trade-offs Between Precision and Recall:** Both models show different trade-offs between precision and recall. BART has a higher recall than T5, meaning it is better at capturing the content present in the reference summaries. However, its precision is also higher, indicating that when BART includes content in its summaries, it is likely to be relevant.
- **Contextual and Semantic Understanding:** The BERT scores, which reflect the contextual embedding comparison, suggest that both models are relatively good at capturing the semantic content of the source material. BART has a slight edge, indicating its summaries might be contextually richer or more nuanced.
- **Fluency and Paraphrasing Abilities:** The METEOR metric, where BART outperforms T5 significantly, suggests that BART is not only capturing the exact words but may also be better at paraphrasing and maintaining the meaning, even when the exact wording differs from the reference.
- **Cosine Similarity and Embedding Space:** Cosine similarity is based on the angle between vectors in high-dimensional space. The higher score for BART indicates that, on average, its summaries are directionally closer to the reference summaries in the embedding space, hinting at better capturing of overall themes and content.
- **Potential for Improvement:** Despite BART's relatively better performance, the BLEU scores for both models are quite low. BLEU is often criticized for not capturing the quality of content that does not match n-grams exactly, but these scores do suggest there might be room for improvement in terms of the model's ability to generate human-like, nuanced summaries.
- **Model Selection for Deployment:** In practical terms, these insights can inform decisions about which model to deploy for a text summarization task. BART appears to be the more robust choice overall, but considerations such as computational resources, domain specificity, and the importance of different types of errors (e.g., favoring false positives over false negatives) would also influence this decision.

Comparison with Baselines

- **Extractive vs. Abstractive Summarization:** Traditional extractive summarization methods select whole sentences or phrases from the source text to form a summary. Both T5 and BART are abstractive models, meaning they generate new text that is not necessarily present in the source, which can lead to more coherent and concise summaries that resemble human-written abstracts.
- **Rule-Based Systems:** Earlier summarization systems often relied on rule-based approaches that followed predefined linguistic patterns. T5 and BART use machine learning to learn from data, which can lead to better generalization and adaptability to different text styles and domains.
- **Statistical Methods:** Statistical summarization techniques, such as those based on term frequency and inverse document frequency (TF-IDF), may lack the nuanced understanding of language that deep learning models possess. T5 and BART can capture complex dependencies and semantic relationships in the text.
- **Earlier Neural Network Models:** Compared to earlier neural network-based models like LSTM or GRU, T5 and BART (which are based on the Transformer architecture) can better capture long-range dependencies in the text, which is critical for summarization tasks.

Impact of Fine-Tuning Data Size

Computational Trade-offs

Larger datasets require more computational resources and time to train, which can increase costs and energy consumption. The complexity of models like T5 and BART means that fine-tuning requires powerful hardware, typically with GPUs or TPUs, which can be a significant investment. The energy required for training large models on extensive datasets has an environmental impact, which is an important consideration in the broader context of AI ethics and sustainability.

Generalizability of Findings

The performance of models fine-tuned on specific datasets may not always generalize to other domains. Models fine-tuned on data from certain linguistic or cultural groups may not perform well on text from different groups due to variations in language use, idioms, and context. Fine-tuning results can be sensitive to the specificities of the dataset, model initialization, and training procedure. This makes reproducibility a challenge, as slight changes can lead to different outcomes.

Ethical Considerations

Implications of Automated Summarization in Healthcare

Summarization can play a role in translating complex medical documents into simpler language that patients can understand, thereby improving patient education and communication. This translation must maintain the nuance and accuracy of the original information. In regions with a shortage of healthcare professionals, automated summarization could help triage patient cases or manage healthcare records more efficiently, potentially leading to better resource allocation.

Bias and Fairness Concerns

- Representative Data: AI models can only be as unbiased as the data they are trained on. If the training data for the summarization model is not representative of the diverse patient populations, the model could generate biased summaries.
- Algorithmic Bias: Even with representative data, algorithms might develop biases based on the patterns they learn. For example, if certain conditions are more commonly documented in one demographic in the training data, the model might overemphasize or underrepresent these conditions in summaries for patients from different demographics.
- Disparate Impact: Biased summaries could have a disparate impact on minority groups or those with less common conditions, leading to unequal levels of care and attention.
- Fairness in Access: There might also be disparities in access to technologies for automated summarization, with well-funded hospitals adopting these tools while underfunded clinics do not have the resources to do so, potentially exacerbating healthcare inequalities.
- Transparency: The methodologies and decision-making processes of AI summarization tools must be transparent, enabling healthcare providers to understand the basis of the generated summaries and to critically assess their validity.

Chapter 7: Conclusion

Summary of Findings

BART outperforms T5 in text summarization across several metrics, indicating better word overlap and semantic alignment with reference summaries. Both models, however, show low BLEU scores, suggesting potential areas for improvement. Larger, diverse datasets enhance model performance but raise concerns over overfitting, increased computational demands, and environmental impact. Ethical considerations, especially in healthcare, necessitate accuracy, privacy, and bias mitigation. There's a critical need for transparency and human oversight in model deployment to ensure fairness and prevent disparities, emphasizing the balance between innovation and ethical responsibility in AI applications.

Limitations

While working on this project, we faced many constraints, primarily in terms of Computational resources. The high computational complexity of models like BART and T5 require access to powerful GPU and TPUs which was a limiting factor. Another limitation was the complexities involved in developing a novel Transformer model, which we tried to do initially but were unsuccessful owing to limited resources and knowledge. These challenges can be addressed by using the latest libraries like JAX, which eases the load of computation and can help develop such models with greater efficiency.

Closing Remarks

This project faced numerous challenges related to managing advanced NLP models and computational constraints, which drove us to think critically and enhance our understanding of text summarization and the implications of deploying AI. We experienced a substantial learning curve, gaining insights into the mechanics of summarization, the importance of high-quality data, and the ethical dimensions of AI applications. Collaboration with stakeholders and the academic community proved invaluable.

Key takeaways include the swift advancements in AI and their transformative potential, highlighting the need for responsible AI that considers ethical implications. Future efforts should focus on enhancing data curation, improving model efficiency, and developing robust evaluation metrics. While AI can boost efficiency and decision-making, human oversight remains crucial for ensuring relevance, accuracy, and ethical deployment.

This project was a profound reminder of the capabilities and limitations of current AI technologies, instilling an appreciation for the nuanced interplay between technology and human oversight, and setting a forward-looking tone for continuous engagement and improvement in the field of AI and text summarization.

Acknowledgment

We would like to thank Professor Arya Farahi for giving us the opportunity to work on this project as the final requirement for his course SDS 384 Scientific Machine Learning. Without his teachings and continuous guidance this project would have never been possible. We would also like to thank the TA Xizewen Han who conducted sessions for PyTorch, which helped us navigate through building Neural network models with ease. Lastly, we would like to thank all our fellow students who helped refine our work and encouraged us to work harder.

Individual Member Contribution:

Team Member	**UT EID**	**Contribution**	**Points (0:No to 100:Full contribution)**
Aditya Ravikant Anekar	ara4462	EDA & Hypothesis testing	100
Jaimik Patel	jp65463	Model training & Validation	100
Srishti Shankar	ss226922	EDA & Hypothesis testing	100
Yug Dave	yd5924	Model training & Validation	100

References

A. R. Lahitani, A. E. Permanasari and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," *2016 4th International Conference on Cyber and IT Service Management*, Bandung, Indonesia, 2016, pp. 1-6, doi: 10.1109/CITSM.2016.7577578.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., & Goharian, N. (2018). A discourse-aware attention model for abstractive summarization of long documents. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. https://doi.org/10.18653/v1/n18-2097

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lavie, Alon & Agarwal, Abhaya. (2007). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. 228-231.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. https://doi.org/10.3115/1073083.1073135

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1-67.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Appendix

Original Long Text Article

“great progress has been made in neonatal care over the last few decades , reflected by improving survival rates and clinical outcomes of preterm infants . despite these advances , 45 years after its first description , bronchopulmonary dysplasia (bpd) remains a major complication of premature birth , causing ongoing morbidity and mortality : it is the most common neonatal chronic lung disease , affecting around 25% to 35% of vlbw neonates (very low birth weight , < 1500 g) , and is associated with increased risk for rehospitalization [3 , 4] , cognitive delay , and neurosensory deficits . initially described by northway et al . in 1967 , the old bpd mainly affected modestly premature newborns suffering from respiratory distress and therefore mechanically ventilated with high levels of supplemental oxygen . with the introduction of surfactant treatment , prenatal maternal use of glucocorticoids , improved nutrition , and ventilator strategies the clinical course and pathology of bpd have changed considerably . unlike the original description , today 's new bpd is mainly regarded as a disruption of distal lung growth [6 , 7] . thus , influenced by both genetic susceptibility [8 , 9] and environmental factors on the immature lung , the pathophysiology is characterized by inflammation , abnormal microvascularization , and impaired alveolarization . alveolar formation of the primitive saccules is a complex process of epithelial morphogenesis , capillary growth , and coordinated extracellular matrix (ecm) remodelling . at this , fibroblast growth factor (fgf) signalling and matrix metalloproteinase (mmp) activity play eminent roles . some mmps are upregulated in inflammatory environment and yet are involved in pulmonary host defense . there is evidence for some mmp isoforms being important determinants for alveolarization , especially mmp-2 , -9 , and -16 : mmp-2 deficient mice showed fewer and larger alveoli with thinner interstitial tissue . hadchouel et al . demonstrated an increase of mmp-16 activity during the alveolar stage and moreover found two snps within the mmp-16 gene being associated with lower tracheal mmp-2 and -16 activity and to protect from bpd . prospecting further potential biomarkers for bpd , also mmp-9 shows some promise ; for example , harijith et al . highlighted a mmp-9-dependent lung injury pathway in an ifn-mediated animal model of bpd . mice with a partial mmp-9 deficiency showed a reversal of ifn-induced lung injury during hyperoxia . mmps , particularly mmp-2 and -9 , activate fibroblast growth factors (fgfs) by cleavage in the ecm , especially during angiogenesis . in turn activated fgfs upregulate mmp expression . fgfs are secreted glycoproteins involved in interactions between epithelium and mesenchyme regulating cell migration and proliferation in embryonic development , especially in fetal pulmogenesis [16 , 17] . their signalling depends on membrane - located receptors (fgfrs) with a tyrosine kinase domain , encoded by four different genes (fgfr 14) [1820] . they are all translated in developing lungs and are suggested to play major roles in modifying distal lung patterns during alveolarization ; for example , fgfr-3-fgfr-4 double - knockout mice show no alveolarization . it has been assumed that heritable determinants contribute significantly to both bpd [8 , 23] and rds . on this account , we were interested in identifying genetic risk factors in a caucasian population of premature newborn with bpd and rds . we genotyped 27 polymorphisms within fourteen candidate genes for bpd : mmp-1 , -2 , -9 , -12 , and -16 , fgf receptors 2 and 4 , fgf-2 , -3 , -4 , -7 , and -18 , signal - regulatory protein (sirpa) , and thyroid transcription factor-1 (ttf-1) . we also included sirpa because of the known effect on surfactant proteins and inhibition of macrophages , as well as ttf-1 due to its effect on lung differentiation . we recruited preterm neonates (28 weeks of gestation)

born between january 1996 and september 2010 at the centre for pediatrics and adolescent medicine , university hospital freiburg , germany . twins and siblings were excluded from the study as were children with chromosomal aberrations , congenital heart defects , or other major congenital malformations . dna was collected by buccal swabs or by routine blood sampling , between 2 weeks up to 2 years of age . this included gestational week , number of days with supplemental oxygen , need of mechanical ventilation and positive airway pressure , and need of surfactant therapy . as described previously , the subdivision of our bpd study population was based on the analysis by lavoie et al . about the heritability of bpd according to the consensus defined by the national institute of health : the bpd population included all infants with moderate and severe bpd , that is , supplemental oxygen for at least 28 days plus need of oxygen and/or positive pressure at 36 weeks of gestation , whereas the control population consisted of all preterm neonates with no or mild bpd . recruiting neonates for the rds population was targeted on severe cases of respiratory distress by including only newborns depending on surfactant within the first 24 hours after birth (see supplementary material available online at http://dx.doi.org/10.1155/2013/932356) . at our neonatal intensive care unit (nicu) the following approach has been applied regarding the treatment with surfactant : avoiding of intubation independent of the gestational week . therefore , even very premature infants are only intubated if they show failure of ventilation and/or need of supplemental oxygen above 40% . once they required intubation during the immediate postnatal period , they receive surfactant within 2 hours . we included a minority of polymorphisms that already had been tested for other pathologies (rs1799750 in mmp1 and rs2276109 and rs652438 in mmp12) . for pcr reactions , genomic dna was initially denatured at 94c for 5 minutes and underwent 3540 cycles of denaturation (94c for 30 seconds) , annealing (1 minute , corresponding temperatures displayed in table 2) , extension reaction (72c for 1 minute) , and a final extension step at 72c for 8 minutes . in table 2 some primers contain intended single nucleotide mismatches (mutagenic primers) to create sites for restriction enzymes . accuracy of the rflp was confirmed by sequencing via dideoxy chain termination method , respectively , three controls (homozygous wildtype , heterozygous , and homozygous mutation) for each polymorphism using the big dye terminator cycle sequencing kit on an abi 310 sequencer (applied biosystems) . genotyping data of our case - control populations were analysed by using armitage 's trend test (att) for possible association with bpd and rds as specified previously . moreover att was used to calculate hardy weinberg equilibrium (hwe) for each polymorphism . the collection of blood / buccal swabs and the experimental procedures were approved by the ethical committee of the university of freiburg . parents were given written and verbal information about the study and a statement of informed consent was signed by the parents of all enrolled children . the results of the 27 studied polymorphisms (table 1) for association with bronchopulmonary dysplasia and neonatal respiratory distress are specified in table 3 (bpd) and table 4 (rds) . among the 11 genotyped polymorphisms in different mmp genes (see table 1) there was no bpd - associated polymorphism (table 3) but two polymorphisms associated ($p < 0.05$) with rds (rs20544 in mmp-9 : $p = 0.033$; rs652438 in mmp-12 : $p = 0.047$, see table 4) . both snps show no significant deviation from hardy - weinberg equilibrium , neither in the control nor in the case population . analysis of rs20544 (c / t) identifies the t allele as protective against respiratory distress . for the genotyping results of the amino acid substitution rs652438 (a / g , asn357ser) the complete absence of the g / g homozygous genotype in the respiratory distress case population must be taken in account . the other mmp - snps showed no association , inclusively rs2664352 in mmp16 , that had been associated with protection from bpd . the fgfr-4 snp rs1966265 , located in the exon region and causing an amino acid substitution of isoleucine (ile) for valine (val) is associated with both bpd ($p = 0.023$) and rds ($p = 0.003$) . here the a / a genotype (ile) could be identified as protective allele variant

against our studied lung diseases . the association results from significant differences in allele frequencies : in both bpd and rds analysis the g allele is more frequent in the disease populations (see tables 3 and 4) . the other snps in the fgfr genes showed no association with neither bpd nor respiratory distress . whereas no association could be detected between the eight fgf - snps and bpd , rs10796856 in fgf-3 and rs4316697 in fgf-7 showed associations with rds . correspondent p values are p = 0.036 (rs10796856) and p = 0.044 (rs4316697) , and no deviations from hardy - weinberg equilibrium were detected (see table 4) . the four snps in sirpa and ttf-1 showed no association with neither bpd nor rds . analysis of ttf-1 rs999460 unfolds deviation from hardy - weinberg equilibrium in both case and one control populations in our caucasian population (see tables 3 and 4) . the aim of this study has been to identify genetic risk factors in an ethnically homogenous caucasian population . genetic contribution to bpd is suggested on the basis of twin studies demonstrating that at least half of the susceptibility is hereditary [8 , 9 , 23] . additionally , lavoie et al . could differentiate in their study that mild bpd (according to the national institute of child health and human development consensus definition) had been mainly attributable to shared environmental factors whereas moderate or severe bpd had been attributable to genetic influence . following these findings , we defined our control population as neonates with no bpd or mild bpd , whereas our bpd population included neonates with moderate or severe bpd . furthermore , we recruited only preterm neonates 28 weeks of gestational age for the bpd population to avoid false associations based on the fact that bpd hardly develops in newborn older than 30 weeks of gestational age . in contrast to bpd , the results of twin studies on rds susceptibility showed mostly contradictory results [24 , 3235] . a twin study by levit et al . with 332 twin pairs of a heterogeneous population has been the first one to include and assess the influence of several independent covariates , revealing that 50% of the variance to rds susceptibility is hereditary . given these lines of evidence for genetic contribution , we have chosen the candidate - gene approach for our association study based on the hypothesis that genes fundamental in lung organogenesis and alveolar remodelling , that is , mmp and fgf , determine susceptibility to bpd and rds . known genetic risk factors for rds are mostly allelic polymorphisms of the genes encoding surfactant proteins sp - a1 , sp - a2 , and sp - b . anyhow , other determinants than components of the surfactant system might also affect the liability to rds . genes encoding for growth factors or enzymes that account for alveolarization through proper secondary septation and extracellular remodeling might affect the gas - exchange and therefore aggravate respiratory distress at birth . supposed genetic risk factors for bpd are mostly genes encoding components of innate immunity and antigen - presentation , cytokines , antioxidant defences , and angiogenic growth factors such as : mannose - binding lectin (mbl2) , tumor necrosis factor - alpha (tnf-) [28 , 38] , human leucocyte antigen (hla)-a , -b , and -c alleles , glutathione - s - transferase - p1 , and vascular endothelial growth factor (vegf) . some years ago , two mmp-16 gene polymorphisms were demonstrated to protect from bpd and moreover to be associated with lower tracheal mmp-2 and -16 levels . matrix metalloproteinases are a family of zinc - dependent endopeptidases , and they degrade extracellular components and play a crucial role in lung development , especially during alveolarization . particularly mmp-2 and -9 (so - called gelatinases a and b) seem to be relevant in extracellular remodeling and even pulmonary host defense . they degrade type iv collagen , fibronectin , elastin , and denatured collagen (gelatin) . mmp-2 deficient mice show an abnormal saccular development with larger and simplified alveoles . in line with this finding , newborns developing bpd showed low mmp-2 tracheal levels at birth [41 , 42] . recently , mmp-9 could be identified as a pathogenic key mediator in a murine model of bpd . on the other hand , increased tracheal levels of mmp-9 early after birth have been associated with resolving rds , suggesting that increase in mmp-9-activity is a physiologic repair response . demonstrated that increased mmp-9 activity in neonatal

lungs early after birth correlated with resolving respiratory distress syndrome , demonstrating a likely role of mmp-9 in pulmonary host defense . in our study we identified an snp (rs20544) in the mmp-9 gene to be associated (p = 0.033) with rds , but not bpd . respiratory distress syndrome has been defined as need of surfactant (see supplementary material) . on one hand , ethnically homogenous populations like our caucasian population are favourable to detect possible pathogenetic determinants , but one must bear in mind that the size of our rds population is limited and the total numbers of neonates studied for each polymorphism vary slightly according to the recruiting time point . furthermore , association studies on rds are prone to confounding factors . other pulmonary conditions such as a transient tachypnea provoked by wet lung syndrome or pulmonary infection might mimic respiratory distress syndrome caused by surfactant deficiency and thereby hamper the results of our study . in our study , we included mmp-16 polymorphisms that had been associated with bpd in a french population (rs2664352) . in our population rs2664352 up to now , the role of fgf3 has been mainly studied in cancer diseases , that is , lung cancer , but its exact role in pulmogenesis remains elusive . there is evidence for fgf-3 upregulation to be associated with alveolar type 2 cell hyperplasia and downregulation to be associated with an excessive recruitment of free alveolar macrophages which might lead to symptoms of respiratory distress . furthermore it has been shown that fgf-3 stimulates the secretion of mmp-2 and -9 propeptides in vitro . fgfr-4 polymorphism rs1966265 showed association with both respiratory distress (p = 0.003) and bronchopulmonary dysplasia (p = 0.023) . the a / a genotype (encoding for isoleucin instead of valine) has been protective in our association study . the exact - test showed no deviation from hardy weinberg equilibrium for this snp in both case and control populations , suggesting that the association does not result from population admixture or genotyping errors . fgfr-1 to fgfr-4 are expressed in the lung and fgfr-3 and -4 signalling , in particular , appears to be fundamental in alveolar formation . weinstein et al . demonstrated that mice deficient in both fgfr-3 and -4 show a completely blocked alveolarization and fail to show any formation of secondary septae , whereas solely fgfr-4(/) animals exhibit no significant abnormalities , revealing a cooperative effect of fgfr-3 and -4 in lung development . hyperoxia - exposed (fio2 0.85) mice show a bpd - like lung pattern of enlarged airspaces and furthermore a reduced expression of fgfr-3 and -4 , suggesting a pathogenic role in arrested lung development . replicated these results in fgfr-3 and -4 deficient mice and demonstrated in addition that fgfr-3/-4 signaling contributes to excessive elastin production and its alveolar accumulation , which is another typical feature of bpd . but these abnormalities have not been due to fibroblast defects but due to increased expression of paracrine factors of alveolar type 2 cell (at2) . if a reduction in fgfr-3 and -4 expression affects distal lung development , a functionally significant polymorphism within the correspondent gene possibly alters the susceptibility to alveolar disease such as bpd and rds . showed that there is a peak of fgfr-4 expression at the day of birth , when respiratory distress syndrome occurs . false - positive results can only be excluded by replications in other study populations . in conclusion , we describe five snps in mmp-9 , mmp-12 , fgfr-4 , fgf-3 , and fgf-7 that are associated (p < 0.05) in our caucasian population with respiratory distress syndrome of the newborn , defined as surfactant application within the first 24 hours after birth . among these polymorphisms one polymorphism in fgfr-4 (rs1966265) is additionally associated with bronchopulmonary dysplasia , demonstrating its possible role in the pathogenesis of newborn lung diseases on grounds of pulmonal immaturity .”