Master Thesis

IU University of Applied Sciences

Study program: Master of Science in Data Science

Enhancing Transparency in Medical Text Transcription Classification: Assessing BERT Pre-trained Language Model Performance and Decision-Making Using Explainable AI (XAI) Techniques

Kaushik Puttaswamy

Enrolment number: 321150196

Meyerheimstr 4

10439 Berlin

Supervisor: Prof. Dr. Thomas Bolz

Date of submission: 06th August 2024

Abstract

This thesis investigates the efficacy and transparency of BERT-based language models for classifying medical text transcriptions, with the intention to meet an existing need for accurate and easily understandable AI in the field of medicine. Conventional machine learning models often lack the essential comprehension of context and transparency needed in medicinal environments. Despite the high accuracy, advanced NLP models like BERT, BioBERT, ClinicalBERT, and RoBERTa have limitations due to their "black box". This work utilizes XAI approaches, specifically SHAP, to enhance the interpretability of models and boost confidence among healthcare professionals. The research utilized a dataset of 4,999 medical text transcription samples to evaluate the traditional algorithms along with the advanced BERT model. The results heavily emphasized the BioBERT's high accuracy and highly favored recall. The SHAP significantly contributed to determining specific characteristics of the model's decisions, thereby aiding in explainability and error identification. As a result, the study will concern itself with several important implications, including clinical decision-making, trust in AI systems, clinical practice, and general healthcare delivery systems. Despite the restrictions of the data set and ethical questionnaires, this research lays down a detailed approach to enhancing the use of AI in medical informatics. This is a way of providing useful information to increase patients' benefits without compromising the accuracy required for interpretation.

Keywords: BERT, XAI, SHAP, medical text transcription, medical informatics, NLP

Table of Contents

List of Figures

List of Tables

List of Outputs

Table of Abbreviations

EHRs	Electronic Health Records
BERT	Bidirectional Encoder Representations from Transformers
XAI	Explainable Artificial Intelligence
ML	Machine Learning
NLP	Natural Language Processing
SHAP	Shapley Additive Explanations
BioBERT	Bidirectional Encoder Representations from Transformers for Biomedical Text Mining
ClinicalBERT	Clinical Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly optimized BERT approach
AI	Artificial Intelligence
SVM	Support Vector Machines
TF-IDF	Term Frequency-Inverse Document Frequency
BOW	Bag-of-Words
BON-gram	Bag-Of-Ngrams
IEEE	Institute of Electrical and Electronics Engineers
NPJ	Nature Partner Journals
CNNs	Convolutional Neural Networks
RNNs	Recurrent Neural Networks
SML	Supervised Machine Learning
RB	Rule-Based
NB	Naïve Bayes
ANN	Artificial Neural Networks
DL	Deep Learning
DT	Decision Trees
RF	Random Forest
LR	Linear Regression
BN	Bayesian Network

ICD-9-CM	International Classification of Diseases, Ninth Revision, Clinical Modification
VA	Veterans Affairs
CAM	Complementary and Alternative Medicine
BIGRU	Bidirectional Gated Recurrent Unit
TCM	Traditional Chinese Medicine
CCKS	China Conference on Knowledge Graph and Semantic Computing
HIT	Health Information Technology
FDA	Food and Drug Administration
MAUDE	Manufacturer and User Facility Device Experience
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative
ROC	Receiver Operating Characteristic
AUCROC	Area Under the Receiver Operating Characteristic
kNN	kNearest Neighbor
LSTM	Long Short-Term Memory
NER	Named Entity Recognition
GPT	Generative Pre-trained Transformer
MIMIC-III	Medical Information Mart for Intensive Care III
Med-BERT	Medical Bidirectional Encoder Representations from Transformers
UMLS-BERT	Unified Medical Language System Bidirectional Encoder Representations from Transformers
UMLS	Unified Medical Language System
CLS	Classification
SEP	Separator
MLM	Masked Language Modeling
NSP	Next Sentence Prediction
ALBERT	A Lite BERT

PSO	Particle Swarm Optimization
BEHRT	Bidirectional Encoder Representations from Transformers for Health Records
G-BERT	Graph-Augmented BERT
GNNs	Graph Neural Networks
PMC	PubMed Central
RE	Relation Extraction
DistilBERT	Distilled Bidirectional Encoder Representations from Transformers
XLM-RoBERTa	Cross-lingual Language Model Robustly optimized BERT approach
BLUE	Biological Language Understanding Evaluation
RETAIN	REverse Time AttentIoN Model
GAM	Generalized Additive Model
LIME	Local Interpretable Model-agnostic Explanations
DeepLIFT	Deep Learning Important FeaTures
MNIST	Modified National Institute of Standards and Technology
CRank	Contextual Rank
ProdLDA	Product of Experts Latent Dirichlet Allocation
NeuralLDA	Neural Latent Dirichlet Allocation
ICU	Intensive Care Unit
RETAIN-Vis	RETAIN Visualization
Grad-CAM	Gradient-weighted Class Activation Mapping
TCAV	Testing with Concept Activation Vectors
t-SNE	t-Distributed Stochastic Neighbor Embedding
PCA	Principal Component Analysis
GDPR	General Data Protection Regulation
AIA	Artificial Intelligence Act
EDA	Exploratory Data Analysis
CSV	Comma-Separated Values
SOAP	Subjective Objective Assessment Plan

ENT	Ear, Nose, and Throat
SMOTE	Synthetic Minority Over-sampling Technique
CUDA	Compute Unified Device Architecture
GPU	Graphics Processing Unit
AdamW	Adam Weight Decay
CV	Cross-Validation
AIMS	Assessment and Information Management System

Chapter 1: Introduction

The field of medical text transcription is crucial in providing quality patient care as well as timely and sound clinical decisions in the constantly evolving health care sector, especially with the increased use of electronic health records (EHRs). The outstanding results in medical text classification have been achieved due to the new NLP tools. Notably, the latest generation of such tools is called BERT (Bidirectional Encoder Representations from Transformers), which has improved the processing of unstructured medical data due to its enhanced contextual orientation. Nevertheless, even some of them use sophisticated machine learning algorithms and encounter interpretability and transparency issues, which are crucial for clinicians who often need to explain their decisions. In response to such challenges, many works, generally known as Explainable AI (XAI), have been identified as essential for helping in the process (Talebi et al., 2024, pp. 2-3). XAI goes a long way in offering better conceptuality and comprehensibility of AI structures, enhancing trust, and achieving the best of medical outcomes ethically.

1.1 Research Background

In the modern world of medicine and its rapid advancement, which results in an increased number of medical records and an exponential increase in the need for medical text transcription data analysis, the classification problem is vital to providing accurate and efficient record-keeping and decision-making for the proper treatment of a patient. In the recent past, there has been improvement in the methods of text classification, bearing in mind the new developments in Natural Language Processing (NLP) and Machine Learning (ML). Regarding these, Bidirectional Encoder Representations from Transformers (BERT) has grown into an outstanding technique that boosts the medical text transcription system effectively. With BERT's bidirectional approach, based on context, it was possible to observe a further enhancement compared to previous models, which underwent numerous difficulties in interpreting medical language and its terms (Ngai & Rudzicz, 2022, p.1). However, the incorporation of BERT as well as other state-of-the art NLP models is a huge concern in clinical practices in terms of the specific model's interpretability. Owing to the fact that deep learning models such as BERT are constantly relied on for diagnosing textual classification for medical use, it is extremely important to understand the decision-making process of these models. To handle these issues, we must use XAI methods like SHAP (SHapley Additive Explanations) to explain how a model arrives at a particular decision. This transparency is required to convince doctors and other healthcare providers that these AI-based systems are augmenting instead of masking their own clinical decision-making (Tjoa & Guan, 2021, p.4793).

The type of research presented in this study revolves around evaluating the effectiveness of, as well as integrating, the different BERT models, such as BioBERT, ClinicalBERT and RoBERTa, with XAI solutions for interpretability reinforcement. The limitations in existing literature have not been addressed, restrictions in medical text transcription have not been improved, and the explainability of

models has not been properly addressed until now; therefore, the objective of this study is to compare these models with traditional machine learning algorithms and use SHAP for explainability. Hence, the findings of this study should enhance knowledge on how resources in NLP technology may be proficiently implemented in healthcare, thus improving the patients' standard of care and the clinicians' decision-making.

1.2 Research Field

This research field focuses on AI, specifically natural language processing and healthcare informatics, with a specific emphasis on the accuracy, transparency, and usability of medical text transcription, as well as the classification of medical texts. Development from rule-based and supervised learning systems to complex deep learning models, especially transformer-based models like BERT, has significantly advanced the field. BioBERT, ClinicalBERT, and RoBERTa, along with the original BERT, have proved ultra-receptive in extracting contextual information along with specific domain terminologies that enhance the accuracy of medical text classification tasks like disease risk prediction, sentiment analysis, and protocol assignment (Chen et al., 2022, p. 4,7).

However, the clinical application of deep learning models has incurred challenges resulting from the black box nature of the models that currently exist, hence the inclusion of XAI techniques that make the models explainable in clinical settings. Such methods as SHAP are used to explain the decision-making processes of AI, making it possible for healthcare professionals to understand the AI-generated prediction and build trust with the AI (Chaddad et al., 2023, pp. 2-3). This combination of the latest NLP techniques with XAI seeks to provide a way of matching intelligent systems with their application in the health sector, especially by ensuring that the existing AI facilitates can be utilized efficiently in these fields.

The research field deals with the problems associated with medical language as it is somewhat ambiguous, variable, and unstructured. To deal with the issues linked with clinical data, the researchers put effort into making the medical text transcription as accurate as it can be in relation to recently developed AI-based predictions. This area focuses on ethical approaches to AI utilization, specifically when it comes to data protection and model bias in relation to the application of AI solutions in delicate healthcare settings. In total, this research field can be characterized as a synergy of leading-edge science and real-world application, the main objective of which is to enhance the AI systems' efficiency and accountability for quantitatively increasing positive health outcomes and qualitatively improving individual-centered patient care toward the realization of the principal concept of precision medicine.

1.3 Problem Statement

Medical informatics is a constantly and rapidly developing area, which is why proper text transcription is crucial in improving the quality of patient treatment as well as in optimizing clinical activities. Although there is some progress in advances in NLP and ML, the existing systems have some challenges when it comes to attaining high accuracy and interpretability. Typical approaches that are based on machine learning face difficulties in understanding the key challenges that arise from medical language use, including the level of ambiguity, variation in the terms used, and the use of unstructured data in clinical work (Ellis et al., 2022, p. 2,4). These difficulties can cause misunderstandings in the classification of clinical texts and hinder the overall application of such AI techniques.

Some of these problems can be solved with the help of new pre-trained language models, such as BERT, which increase the efficiency of medical text classification. Despite the enhanced performances of models such as BioBERT, Clinical BERT and RoBERTa against traditional approaches, these models subsequently entail non-interpretable mechanisms. This kind of opacity programming makes it hard for the healthcare staff to comprehend and therefore have faith in the recommendations made by the AI, hence hindering the use of AI in clinical practice. One way of closing this gap is through the leveraging of XAI methods such as SHAP (SHapley Additive exPlanations) to fill the gap and increase the model's interpretability (Lundberg & Lee, 2017, p. 3).

Consequently, the issue that this thesis seeks to solve is that of improving classification systems for medical text transcription while also improving the interpretability of the output. This includes not only comparing the performance of the new BERT-based models but also how these demands facilitate the incorporation of XAI for explaining the decision-making of the novel models. In this way, the existing research will target the objectives of improving health care based on accurate AI and developing better and more effective algorithms that could be used in the practice of medicine.

1.4 Aims and Objectives

a) Aim

1) To Evaluate the Performance of BERT Models in Medical Text Classification:

- Evaluate the performance of different BERT-based models, such as BioBERT, ClinicalBERT, and RoBERTa, in classifying medical text transcription.

2) To Enhance Transparency and Interpretability of Medical Text Classification Models:

- Integrate XAI approaches, notably SHAP, into BERT models to increase predicted interpretability and model decision-making.

3) To Address Challenges in Medical Text Transcription Through Advanced NLP Techniques:

- Utilize advanced NLP algorithms to address ambiguity, terminology variability, and unstructured data in medical text categorization.

b) Objectives

1) Data Collection and Preprocessing:

- Collect and preprocess a complete dataset of medical text transcriptions to ensure quality and relevance for model training and evaluation.
- Use data normalization, tokenization, and encoding techniques to prepare datasets for BERT-based models.

2) Model Implementation and Comparison:

- Implement traditional ML algorithms (e.g., SVM, Random Forest, XGBoost and logistic regression) for baseline comparison in medical text classification.
- Implement and fine-tune BERT-based models, including BioBERT, ClinicalBERT, and RoBERTa, for enhanced performance in medical text classification.

3) Performance Evaluation:

- Evaluate and compare the performance of traditional ML models and BERT-based models using metrics such as precision, recall, F1 score, and accuracy.
- Analyze the impact of text representation techniques (e.g., Bag-of-Words vs Bag-of-Ngrams vs TF-IDF) on model performance.

4) Integration of Explainable AI:

- Integrate SHAP with BERT-based models to provide insights into feature importance and model decision-making processes.
- Assess the effectiveness of SHAP in improving model transparency and interpretability, and its utility in identifying and addressing biases in predictions.

5) Analysis of Model Transparency and Trust:

- Investigate how SHAP visualizations contribute to trust and understanding of model predictions by medical professionals.
- Evaluate the ability of SHAP to assist in error analysis and debugging of BERT-based models.

6) Exploration of Ethical and Practical Considerations:

- Address ethical considerations related to data privacy, fairness, and accountability in the deployment of AI-driven medical text classification systems.
- Examine practical challenges in scaling and implementing BERT-based models with XAI techniques in real-world clinical settings.

1.5 Research Questions

a) Performance Evaluation of BERT Models:

- How do BERT-based models (BioBERT, ClinicalBERT, and RoBERTa) compared to traditional machine learning algorithms (e.g., SVM, Random Forest, XGBoost and logistic regression) in terms of accuracy, precision, recall, and F1 score for medical text classification?

- What is the impact of different text representation techniques (e.g., Bag-of-Words vs Bag-of-Ngrams vs TF-IDF) on the performance of BERT-based models in medical text classification?

b) Transparency and Interpretability of BERT Models:

- How does the integration of SHAP with BERT-based models affect the interpretability and transparency of predictions in medical text classification tasks?
- What insights can SHAP provide regarding feature importance and decision-making processes of BERT models, and how does this enhance trust and understanding among medical professionals?

c) Challenges in Medical Text Transcription:

- What are the primary challenges associated with medical text transcription, such as ambiguity, terminology variability, and unstructured data, and how effectively do BERT-based models address these challenges compared to traditional machine learning approaches?
- How do advanced NLP techniques, including BERT and domain-specific models like BioBERT, overcome issues related to the complexity and variability of medical language?

d) Integration of Explainable AI:

- To what extent does the application of SHAP improve the ability of BERT-based models to provide explanations for their predictions, and how does this integration impact the overall performance and reliability of these models?
- How effective is SHAP in assisting with error analysis and debugging of BERT-based models, and what role does it play in identifying and addressing biases in predictions?

e) Ethical and Practical Considerations:

- What ethical considerations, including data privacy, fairness, and accountability, need to be addressed when deploying AI-driven medical text classification systems with BERT and XAI techniques?
- What practical challenges are encountered in scaling and implementing BERT-based models with XAI techniques in real-world clinical settings, and how can these challenges be mitigated?

1.6 Significance

The impact of this research is the possibility of revolutionizing the methods for medical text transcription classification, thus connecting state-of-the-art NLP techniques with real-world clinical needs. This work also aims at comparing the BERT-based models, specifically BioBERT, ClinicalBERT, and RoBERTa, with the traditional machine learning algorithms and also to demonstrate that the pre-trained models are more efficient in dealing with challenges in the medical text. To this end, in addition to the above-mentioned methods, XAI techniques, especially SHAP, have been incorporated to enhance the model's explainability and thus counteract some of the concerns related to the "black box" of deep learning systems. This is because transparency is beneficial in the development of trust among medical workers, in the analysis of errors made, and also in the encouragement of the right

use of AI in the healthcare industry (Fehr et al., 2024, p. 2). Moreover, the research also addresses some general problems of medical text transcription, including ambiguity, terminology variability, and unstructured data, describing how the advanced NLP techniques cope with these issues. Thus, the study also discusses the ethical concerns of data privacy, fairness, and accountability, which are crucial for the proper implementation of AI in the clinical environment. Altogether, this research benefits the field of precision medicine and enhances the quality of patient care by increasing the accuracy and efficiency of medical text classification. Thus, the role of integrating advanced AI technologies with explainability to improve the effectiveness and credibility of AI in healthcare is highlighted.

1.7 Structure of Dissertation

The dissertation starts with an introduction that defines the purpose of the study, which is to improve the level of transparency in medical text transcription classification using BERT pre-trained language models and XAI techniques. It entails background information about the proposed study, the definition of the research problem, the overall purpose and objectives of the research, and questions to be answered. Which moves to the next section of the literature review, which provides a comprehensive assessment of prior studies related to medical text transcription, progress in machine learning, and the combination of BERT and XAI methods. They include emphasizing historical developments, enabling the presentation of perception gaps in the existing knowledge, and addressing ethical issues. Chapters on research methodology also explain how the data are collected, preprocessed, and used to develop the models, including BERT and XAI methods in benchmarking and interpretation. It also discusses validity and reliability and has considerations, especially for the ethical approaches to be taken. The corresponding sections are named research findings and interpretation are discussed here in terms of the usage of all the mentioned methods and the efficiency of the BERT model and XAI techniques in the aspect of transparency enhancement. The conclusion of the dissertation is the summary of the findings, the discussion of the theoretical as well as practical relevance of the work, and the proposal of further research ideas, which form the impact and contributions to the field. Figure 1 below provides flow diagrams of the structure of the dissertation and highlights the logical relationships among the chapters to be included in the study.

Figure 1. Structure of dissertation

Source: Own representation.

Chapter 2: Literature Review

2.1 Introduction

A literature review is a synthesis and critical evaluation of works already published in a particular field of study or subject area. It aims at capturing the existing knowledge, with a focus on the gap that may exist or the latest published research breakthrough or trend (Ramdhani et al., 2014, p. 48). The present literature review seeks to explore the impact brought about by the application of BERT and XAI techniques on the classification of medical text transcription when implemented in health care decisions. It reviews recent work and advancements in extracting information from medical texts that have incorporated BERT, with attention to such challenges as sentiment analysis, disease prognosis, medical image protocol designation, and others. Moreover, the review examines the proficiency of XAI in enhancing the AI-driven systems interpretability and transparency, which is extremely important for raising clinician confidence in clinical application and, consequently, enhancing patient outcomes.

The literature review also compares the effectiveness of BERT and XAI methods in addressing challenges pertaining to the classification of issues specific to medical text transcription, including medical language ambiguity, limited data availability, and interpretability of models. Moreover, this review analyzes several works and applications that demonstrate benefits from such integration, such as an enhancement of BERT's contextual understanding and XAI's explanatory aspect; it also offers an understanding of how these improvements in turn help fine-tune better and more accurate healthcare information systems. In addition, it discusses new directions for further research and the trends focused on the elimination of present limitations and the better inclusion of these technologies in clinical practice.

2.2 Literature Review Process

The process of literature evaluation included searches in the most popular databases and specialized platforms, including Google Scholar, IEEE Xplore, NPJ Digital, Springer, Elsevier, etc. The emphasis was placed on obtaining new information on applying BERT and XAI approaches in the field of healthcare. The review focused on peer-reviewed journals, conference proceedings, and research articles that are concerned with the application of ML in healthcare. The first objective was to review the developments in the classification of medical text transcription and to explore how future advancements in the fields of ML and NLP can contribute to the improvement of the reliability and validity of AI-based decision-making systems in clinical practice.

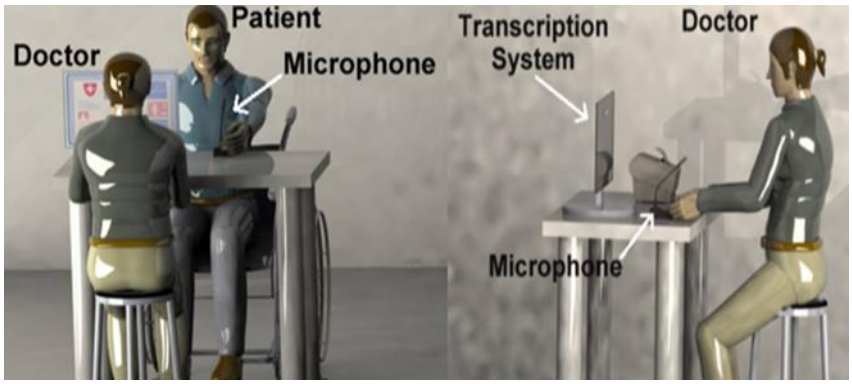
This structure allowed for the systematic coverage of important topics, including medical text transcription classification, and provided a view of the future of AI in medicine. Thus, the review was focused on emphasizing the possibilities of using BERT and XAI for enhancing decision-making in clinical practice by analyzing the previous advancements. The synthesis of findings not only reveals

the progress of technology but also points out the problems and potential developments of the next research project. A special focus is placed on ethical issues, legal requirements, and the role of XAI in improving health care.

2.3 Medical Text Transcription Classification

Medical transcription plays a crucial role in healthcare to transcribe the spoken medical information into digital form, reduce the communication gaps between the providers, and avoid errors that have a direct impact on patient safety and healthcare services (Bhandari, n.d.). Accurate transcription is essential for creating and sustaining comprehensive and accurate EHRs, which are necessary for a patient's medical history, diagnosis, treatment, and outcomes, which are important for continuity of care, clinical decision-making, and legal needs (Pagad et al., 2022, p. 2). Medical text transcription classification is the process of organizing textual data into specific categories, which helps in data organization, analysis, and decision-making for further research and clinical practice regarding patients' demographics, diseases, operations, therapies, and results (L. Yao et al., 2019, p. 32). Despite its importance, the complexity of medical language presents significant challenges. Figure 2 illustrates the medical text transcription process in healthcare operations.

Figure 2. Medical text transcription process



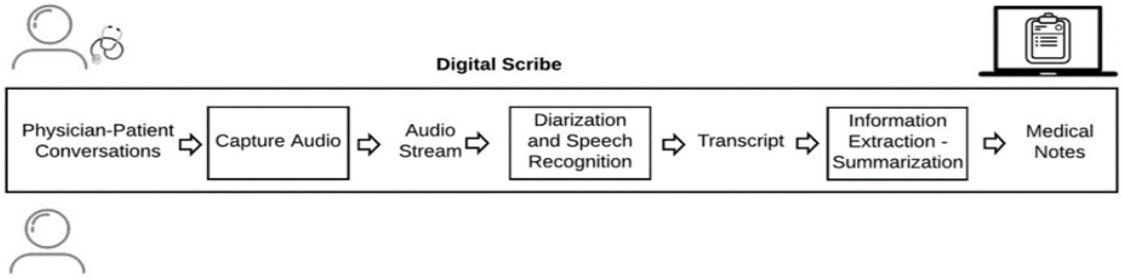
Source: Adapted from: Zhang et al., 2023, p. 2.

2.3.1 Importance and Challenges of Accurate Medical Text Transcription and Classification

Medical text transcription is vital for preserving the patient's record, facilitating communication between the provider and the patient, and assisting in decision-making in patient history documentation, symptoms, and treatment (Almazaydeh et al., 2023, pp. 66-67). As shown in Figure 3, digital scribes that use AI and ML to transcribe notes from the clinician-patient interactions improve the functionality of EHRs. Classification of these transcriptions into appropriate categories also enhances the delivery of health care since it enhances information retrieval, correct billing and management of epidemics and other diseases. (Zhang et al., 2023, pp. 1-2). Medical text transcription and classification face challenges due to the ambiguity and complexity of medical terminology, contextual nuances, and variations in language usage across specialties (Khurana et al., 2023, p. 3718). These problems are solved using the methods of named entity recognition (NER) and relation

extraction, as well as the CNN and RNN deep learning models (Weng et al., 2017, p. 2,11). Other techniques such as domain specific knowledge bases, transfer learning, and semi-supervision are also used to improve the performance of the system and deal with the problem of limited annotated data (J. Yao et al., 2023, p. 2). Additionally, the use of XAI methods, including SHAP, is paramount to ensuring the model's interpretability and reliability in the medical field.

Figure 3. Digital scribe automating medical note tasks from clinician-patient conversations



Source: Quiroz et al., 2019, p. 2.

2.3.2 Evolution of Medical Text Transcription: From Manual Methods to Digital Technologies

The process of medical transcription has changed from writing notes by hand to complex digital systems, which have greatly enhanced the documentation of healthcare. First, physicians employed abbreviations; however, the appearance of medical transcriptionists in the late nineteenth and early twentieth centuries improved both the precision and thoroughness of the records (Cora Garcia et al., 2010, pp. 87-88). Some of the early technologies that were adopted in the 1950s include dictation machines and magnetic tape technology, which improved efficiency (MOS, 2021). The transition to digital systems in the later twentieth century and the incorporation of speech recognition also had a positive effect on the field as documentation and the flow of information between inter-provider collaboration were enhanced, which led to better patient outcomes (Eftekhari, 2024, p. 554).

The field of medical transcription is still dynamic due to the developments in AI, ML, and NLP that enhance the analysis of unstructured medical information (Filipp, 2019, p. 209). These technologies improve the accuracy and efficiency of transcription, which is critical to patient-centered care and precision medicine (Hu et al., 2020, p. 10). The growth of robust NLP tools is an important prerequisite for structuring and extracting data from free text, and the advances in speech recognition technologies have made this tool useful in clinical practice (Barr et al., 2017, p. 6). Future developments will rely on these technologies to transform the text into structured data, which will improve the efficiency of healthcare.

2.4 Machine Learning in Medical Text Transcription Classification

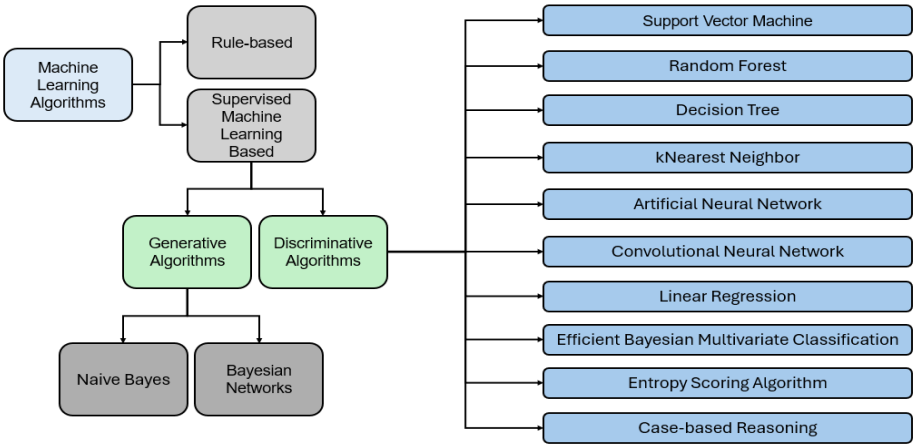
The advancement in technology has seen medical records go digital, and this has resulted in an increase in textual data, hence the need for proper categorization and data mining. Machine learning has been widely used for automating the classification of medical text transcription to improve

accuracy and efficiency. New developments are directed towards tailoring algorithms into medical language, enhancing efficiency and flexibility in different aspects of medicine (Sreekumar & Nizar Banu, 2022, pp. 613-614). This review aims at identifying the progress made, the current issues, and the future research prospects in medical text transcription classification using machine learning.

2.4.1 Traditional Machine Learning Classification Approaches, Recent Studies for Clinical Text Classification, Evaluations Measures and Limitations

a) Traditional ML classification approaches: Conventional methods of ML are widely used in the categorization of medical texts, which helps to improve the automation of the analysis of clinical data. In this context, the most important techniques are Supervised Machine Learning (SML) and Rule-Based (RB). SML algorithms are of two types: generative and discriminative algorithms, which are commonly used. While generative models like Naïve Bayes (NB) estimate the joint probability distribution of features and classes, discriminative models like Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Convolutional Neural Networks (CNN) estimate the conditional probability distribution, with more emphasis on the differences between classes. SVM, in particular, is known to be efficient for both linear and non-linear data, but it has limitations in terms of memory and interpretability (Mujtaba et al., 2019, pp. 507-510). These algorithms are depicted in Figure 4 below in different text classification contexts.

Figure 4. Machine learning algorithms used in different scenarios of text classifications

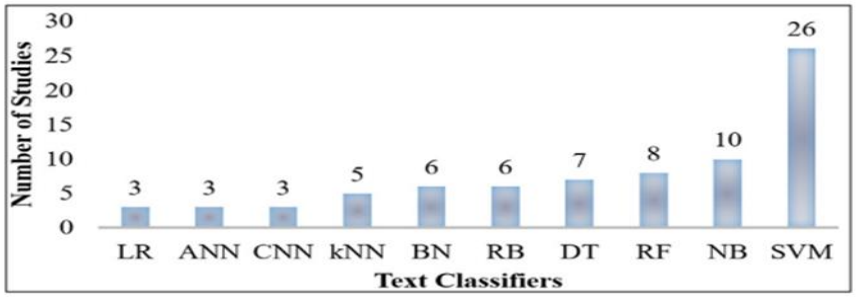


Source: Adopted from: Mujtaba et al., 2019, p. 508.

The Rule-Based systems, which use the rules defined by the domain specialists, provide high accuracy and flexibility in the categorization of the medical text. They are appreciated for their straightforwardness and the possibility of easy debugging, but they depend on expert knowledge, which makes them less applicable and less universal (Deng et al., 2015, p. 1038). Mujtaba et al. (2019, p. 508), pointed out that, as illustrated in Figure 5, SVM is the most employed classifier in the medical text categorization studies in comparison with other classifiers, including NB, DT, and RF. ANN and CNN are used for pattern recognition, while DT and RF are used because of their interpretability.

The choice between the methods depends on the nature of the data, the classification problem, and the required interpretability.

Figure 5. Graph showing the frequency count of text classifiers used in previous studies



Source: Mujtaba et al., 2019, p. 510.

b) Related works: This section briefly describes several current works on clinical text classification with different ML approaches.

Guo et al., (2016, pp. 824-825) developed a multi-label text classification approach for health records, where the authors concentrated on feature selection to solve the problem of training data. They employed a forward search strategy and prediction risk for feature relevance and enhanced the classifier. Their method, which was applied to 1566 clinical records, included feature extraction of TF-IDF and stop and stem word filtering, which allowed for the effective classification of diseases by ICD-9-CM codes.

Ong et al., (2010, pp. 2-4) used a technique to automate the classification of clinical event reports to identify critical incidents such as poor clinical handover and incorrect patient identification. They compared SVM with radial-basis and linear kernels and Naïve Bayes classifiers and got about 80% accuracy with little data. Their approach showed that it is possible to achieve large-scale classification of incidents in the healthcare domain.

Shao et al., (2018, pp. 2-4) conducted a study to compare the word embedding features (doc2vec, word2vec) with the traditional bag of words features in clinical text classification. They also identified that, by utilizing records from the VA on Complementary and Alternative Medicine (CAM), word2vec was superior to BOW-1-gram features, while BOW-2-grams were inconclusive. In their study, they focused on the effects of feature representation on classification.

Qing et al., (2019, pp. 2-8) proposed a neural network approach that uses Bidirectional Gated Recurrent Units (BIGRU) and convolutional layers, with an attention mechanism for medical text classification. In the experiments on traditional chinese medicine and CCKS conference data, their method has achieved promising performance in the classification of medical texts.

Chai et al., (2013, pp. 2-5) used statistical text categorization to identify HIT issues from the FDA and MAUDE databases. They used logistic regression and feature selection techniques and realized that stemming reduced the feature set size to a great extent and improved the recall rate to 0. 989

but reduced the precision rate to 0. 165. In their study, they emphasized the use of statistical classification for the identification of HIT incidents and recommended future work on semi-supervised learning.

c) Evaluations measures: The following measures are employed to evaluate the performance of a medical text categorization model: accuracy, precision, recall, F-measure, specificity, sensitivity, and Area Under the Curve (AUC). These measurements are based on the confusion matrix values: It is made up of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). The common measures used in binary classification problems are precision, recall, F-measure, accuracy, AUC, sensitivity, and specificity. The averaging of precision, recall, and F-measure is usually done by micro- and macro-averaging in multi-class problems. The performance metrics are described as follows:

- Precision: Precision is defined as the ratio of the number of correctly predicted positive medical categories to the total number of positive medical categories that have been predicted. It is also known as a positive predictive value.
- Recall: The number of correctly predicted positive medical categories is divided by the total number of medical categories in the positive class. It is also known as a true positive rate or sensitivity.
- F-measure: The F-measure is the weighted average of precision and recall, yielding a single score that balances both criteria.
- Accuracy: Accuracy is the most utilized performance metric. It is the proportion of accurately predicted medical categories to total medical categories.
- Receiver Operating Characteristic Curve (ROC curve): The ROC curve compares the rate of true positives with false positives.
- Area Under the ROC Curve (AUCROC): The area under the ROC curve measures a classifier's overall performance. A value near to one suggests superior performance.
- Specificity: The proportion of negative instances correctly predicted as negative.
- Micro- and Macro-average of Precision, Recall, and F-measure: Techniques for aggregating precision, recall, and F-measure scores from different classes. Micro-averaging adds individual TP, FP, and FN values, whereas macro-averaging computes the average precision, recall, or F-measure across multiple sets.

These evaluation measures give information about several aspects of a model’s performance, including its ability to correctly classify positive instances, its overall accuracy, and its ability to perform well on multiple classes (Sokolova & Lapalme, 2009, pp. 429-430).

d) Limitation: Linear regression, artificial neural networks, k-nearest neighbor, and other such methods have been the traditional pillars of predictive modeling. However, each of them has its own drawbacks. Linear regression has limitations because it presupposes linearity between the input variables and the output variables, which may not be the case in real-world data sets. It is affected

by outliers and multicollinearity in the data (Sarker, 2021, p. 8). ANNs and CNNs, though very effective, are often accused of being "black boxes" due to their complex internal architecture that hinders the understanding of the decision-making process. They also need large amounts of training data and are prone to overfitting (Brnabic & Hess, 2021, p. 17). Similarly to kNNs, kNNs depend on the chosen value of k and distance measure and can be computationally intensive for large datasets. It also has a problem when dealing with high-dimensional data. While RFs are more interpretable than individual decision trees, they also have the "black box" problem, and their performance depends on the choice of hyperparameters. SVMs are not easy to analyze because the decision boundary is created by a complex mathematical formula, and they are also prone to the selection of the kernel function and hyperparameters (Boateng et al., 2020, p. 344). Bayesian networks require the input of prior probabilities that can be hard to obtain, and the models can be computationally intensive, especially when complex models are used (Brnabic & Hess, 2021, p. 4). The problem with rule-based systems is that it becomes difficult to manage the rules as the number of rules increases, and it may not work well with noisy or incomplete data, while decision trees may overfit, especially if the tree is deep, and may not work well with continuous or high-dimensional data (Ozcan et al., 2022, p. 3291). Naive Bayes classifiers rely on significant assumptions on feature independence, which in real-world datasets may not hold, and they are sensitive to the quality of the training data (Boyko & Boksho, 2020, pp. 5-6).

To overcome the above limitations, advanced strategies such as ensemble methods, deep learning, and XAI have emerged. Combining techniques involves taking several models and making them work in parallel to provide better performance and something that is easier to understand. As a result, the structures of deep learning contain the potential capability for multifaceted data processing, requiring large amounts of computational processing. The concept of XAI is aimed at presenting the decisions made by the AI models in front of different stakeholders, which helps in building trust (Yang et al., 2023, pp. 165-166). These innovations mitigate the issues that are present in classic methods of ML by enhancing interpretability, scale, and reliability in order to make ML applications more trustworthy in their diverse domains.

2.4.2 Need for Transparent and Interpretable Deep Learning Models in Medical Text Classification

Medical text categorization is useful in medical natural language processing for identifying important information from clinical documents, including electronic health records and medical notes. In the past, feature engineering was done manually, but deep learning algorithms have enhanced classification since they handle unstructured data well. The latest developments include the quad-channel hybrid LSTM and the hybrid BiGRU with multihead attention to improve the classification rate and feature extraction (Prabhakar & Won, 2021, pp. 2-3). However, deep learning models have been criticized for their black box nature and interpretability, which is critical in the medical domain where model decisions need to be explained. To deal with these issues, the researchers are working on

developing interpretable deep learning models for medical text categorization. One of them is attention mechanisms that help models focus on the necessary parts of the text and explain the classification results (Qing et al., 2019, p. 2). Furthermore, the integration of rule-based and deep learning with weak supervision can enhance interpretability as well as minimize the necessity of a large amount of manual annotation (Lu et al., 2022, p. 3). These efforts aim to develop models that clinicians can trust and use effectively in decision-making.

2.4.3 Natural Language Processing in Medical Text Analysis

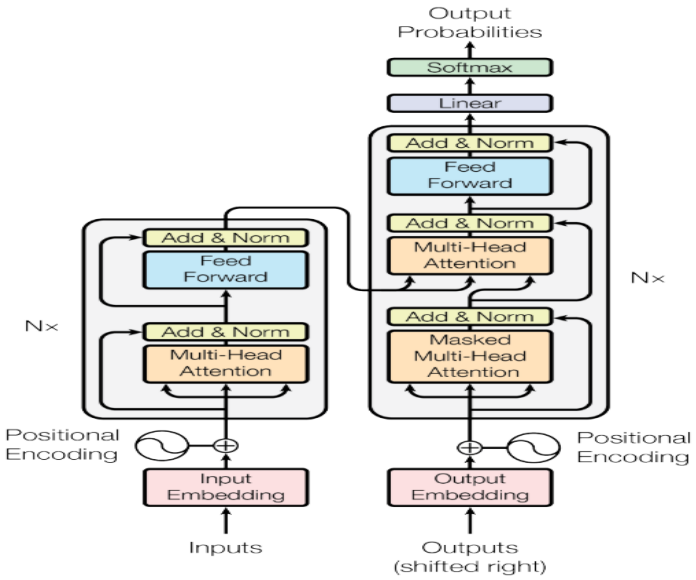
NLP is currently a critical component of medical text analysis, which has altered the healthcare industry by allowing relevant information to be extracted from clinical documentation. Tokenization, NER, and part-of-speech techniques aid in information extraction and coding automation, whereas sentiment analysis and language models such as BERT and GPT aid in clinical decision support and question answering systems (Janowski, 2023, p. 52). The most recent developments include deep learning and transformer-based models such as BioBERT and ClinicalBERT that offer advanced linguistic features that are suitable for biomedical and clinical domains. These models have been further fine-tuned for domain-specific pre-training; for instance, BioBERT trained on PubMed data and ClinicalBERT trained on MIMIC-III notes have shown a remarkable improvement in named entity recognition and question-answering tasks (Huang, Altosaar, et al., 2020, pp. 3-4). The introduction of models like RoBERTa, which improves BERT's pre-training approach, has propelled NLP performance in healthcare to new heights (Talebi et al., 2024, p. 7). Furthermore, emerging models such as Med-BERT and UmlsBERT, which combine medical knowledge from sources such as the UMLS Metathesaurus, represent a significant advancement in the development of accurate and efficient NLP applications in the healthcare domain (Rasmy et al., 2021, p. 5). These advancements illustrate the growing capability of NLP to enhance healthcare delivery through improved text analysis and decision support.

2.5 BERT and Pre-trained Language Models

Language model pre-training is important in NLP, where BERT is an example of the fine-tuning method. Developed by Google researchers in 2018, BERT applies bidirectional training with a masked language model to predict the missing words and the relations between the sentences; thus, it captures more contextual information (Devlin et al., 2019, pp. 4174-4175). Unlike the previous encoder-decoder models, BERT uses the transformer model with self-attention and feed-forward layers. This design involves multi-head self-attention and positional encoding to enable parallel computation and enhance the model’s capacity to address long-distance dependencies (Vaswani et al., 2017, pp. 2-3). This has greatly enhanced the field of NLP and its applications because of the transformer’s capacity to flexibly adjust the importance of words and process information from different representation subspaces. Due to the limitations of sequential processing in RNNs and the large-scale pre-training, BERT and similar models have set new records in many NLP tasks. This

advancement demonstrates the change that pre-training techniques have brought to the field (Paaß & Giesselbach, 2023, pp. 26-27), as illustrated in Figure 6.

Figure 6. The transformer - model architecture

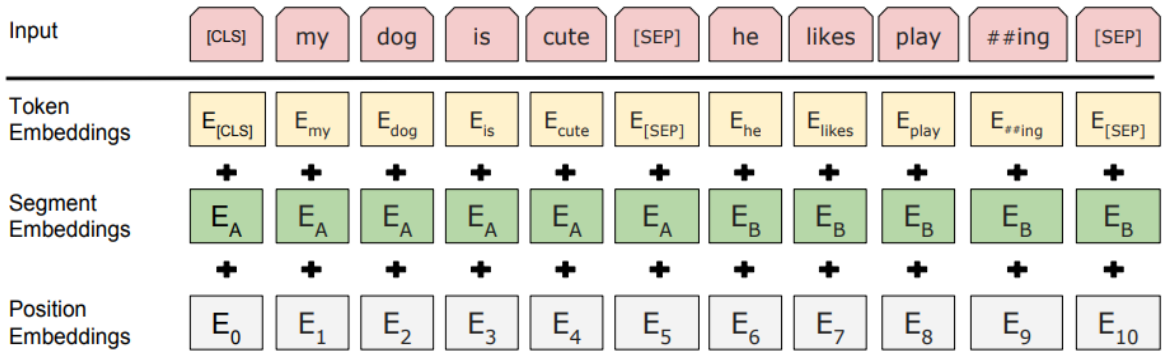


Source: Vaswani et al., 2017, p. 3.

2.5.1 Leveraging Pre-trained Language Models for Text Classification: Working Principles of BERT, Advantages and Developments and Impact

a) Working Principles of BERT: BERT is a natural language processing model that is based on a multi-layer bidirectional transformer encoder. It comes in two variants: The two models are BERT-BASE with 12 layers, 768 hidden units, and 12 attention heads, and BERTLARGE with 24 layers, 1024 hidden units, and 16 attention heads, with BERTLARGE having more parameter capacity. BERT uses WordPiece embeddings with 30,000 subwords and special tokens such as [CLS] for classification and [SEP] for separating sentences. Thus, the input representations include token, segment, and position embeddings to encode phrases and sentence pairs efficiently (Devlin et al., 2019, pp. 4174-4175), as illustrated in Figure 7.

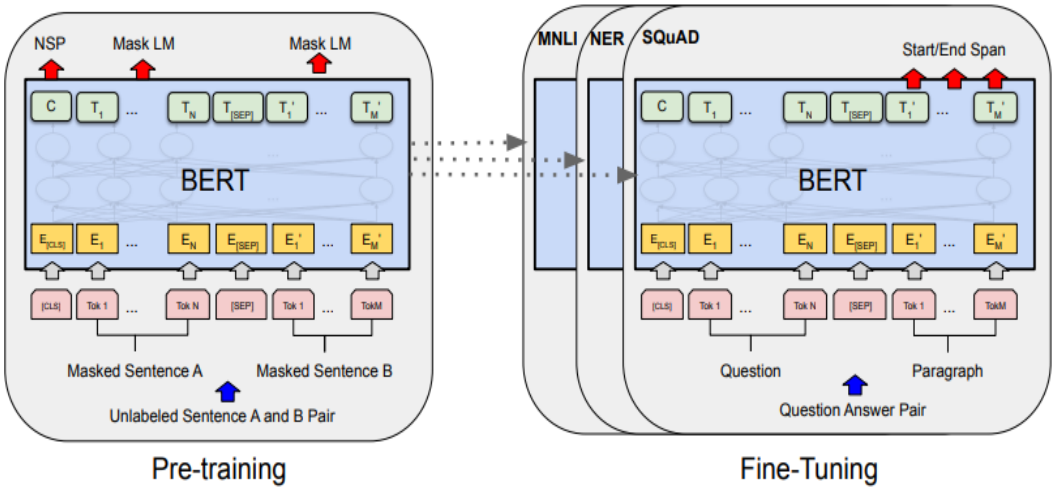
Figure 7. BERT input embeddings overview



Source: Devlin et al., 2019, p. 4175.

The operational model of BERT includes pre-training on large corpora such as BooksCorpus and English Wikipedia using the MLM and NSP techniques. In pre-training, MLM is used to predict the masked tokens in the sequences, while NSP is used to measure the relationship between the sentences. After pre-training, BERT is further trained with task-specific labeled data to achieve high performance in various tasks, such as question answering and text classification. This fine-tuning process uses the transformer's self-attention to address the different task-related inputs and outputs (Devlin et al., 2019, pp. 4172-4173), as shown in Figure 8.

Figure 8. BERT: architecture and fine-tuning overview



Source: Devlin et al., 2019, p. 4173.

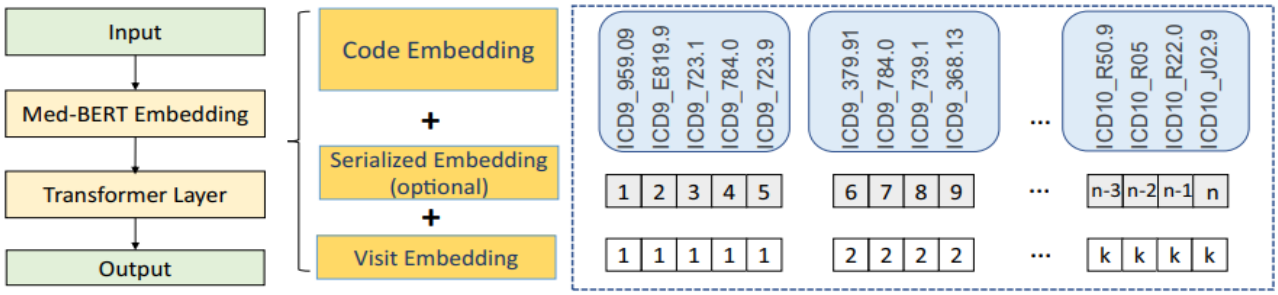
b) Advantages and Developments: State-of-the-art language models like BERT have greatly improved text classification tasks by using a large amount of unlabeled data for pre-training. This pre-training allows models to learn semantic features that are useful for downstream tasks without the need for task-specific annotations; when fine-tuned with scarce labeled data, models pre-trained in this way perform significantly better (Wang et al., 2023, p. 12, 25). These models are effective at dealing with polysemy and ambiguity by using contextualized embeddings and are applicable in many fields and tasks like sentiment analysis and entity recognition (Zhu et al., 2024, pp. 1-3). Pre-trained models make NLP development easier because they can be fine-tuned, and this process is not time-consuming and requires few computational resources. Additional pre-training on domain-specific corpora can improve performance on specific tasks, especially when labeled data is limited (Zhao et al., 2021, p. 3). Further, efficient models such as ALBERT also resolve the issues of computational and memory limitations, especially in resource-limited settings (Lan et al., 2020, p. 2). Thus, pre-trained language models have significantly improved NLP performance; however, continuous research is needed for advancements.

c) BERT's Impact on Medical Text Analysis: Recent breakthroughs in NLP, especially in the application of transformer models such as BERT, have greatly affected the analysis of medical texts. The use has advanced to areas such as medical chatbots and EHR analysis. The medical chatbots with

BERT provide the relevant medical data and appointment scheduling and employ such datasets as MIMIC-III and COVID-19 to guarantee accuracy and compliance with the laws of data protection (Babu & Boddu, 2024, p. 3). Nevertheless, there is still potential to improve their conversational quality and focus on relevant domains.

Med-BERT, which is developed specifically for structured EHR information, has three kinds of embeddings: code, serialization, and visit embeddings to address the challenges of EHRs. This approach assists in disease prediction in addition to providing information on the length of stay by taking into consideration semantics within clinical codes (Rasmy et al., 2021, p. 2,5). Figure 9 illustrates Med-BERT's architecture and embedding layers. However, further research is needed to assess Med-BERT's scalability and generalizability across various healthcare settings.

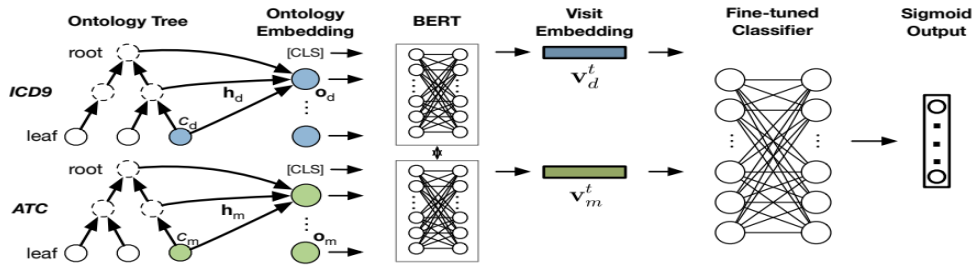
Figure 9. Med-BERT architecture and embedding layers



Source: Rasmy et al., 2021, p. 5.

Another more developed adaptation is called BEHRT, which improves EHR predictive modeling through the use of disease embeddings and positional encodings and proves to be more accurate in disease prediction than traditional approaches (Y. Li et al., 2020, p. 4). Graph-Augmented BERT (G-BERT), which integrates graph neural networks and transformer networks, enhances medical text analysis by learning the hierarchical and relational features from medical ontologies (Shang et al., 2019, pp. 2-3). The G-BERT framework is depicted in Figure 10 below. Ontology embeddings with BERT are incorporated into G-BERT to develop a better contextualized understanding of clinical data.

Figure 10. The framework of G-BERT



Source: Shang et al., 2019, p. 3.

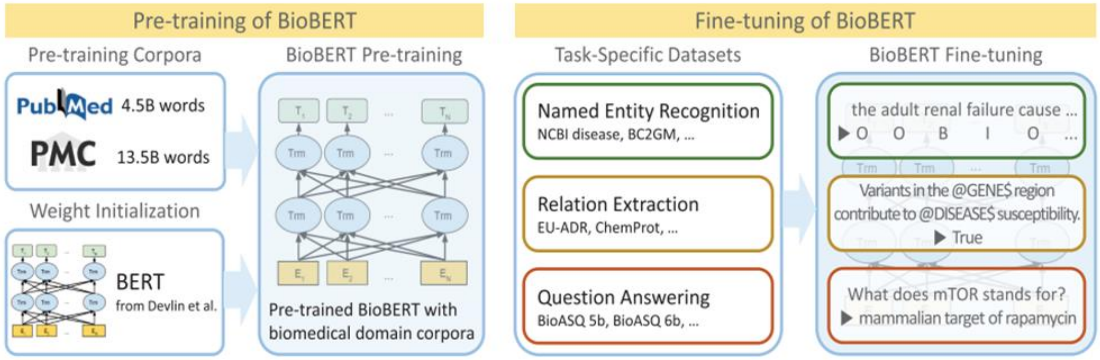
BioBERT is developed specifically for biomedical text mining since it is pre-trained on the PubMed abstract and PMC articles to capture domain-specific terms (Lee et al., 2020, p.1235). The pre-training corpora employed for BioBERT are listed in Table 1, and an overview of BioBERT's pre-training and fine-tuning is shown in Figure 11.

Table 1. BioBERT pre-training on various text corpora

Model	Corpus combination
BERT (Devlin *et al.*, 2019)	Wiki + Books
BioBERT (+PubMed)	Wiki + Books + PubMed
BioBERT (+PMC)	Wiki + Books + PMC
BioBERT (+PubMed + PMC)	Wiki + Books + PubMed + PMC

Source: Lee et al., 2020, p.1236.

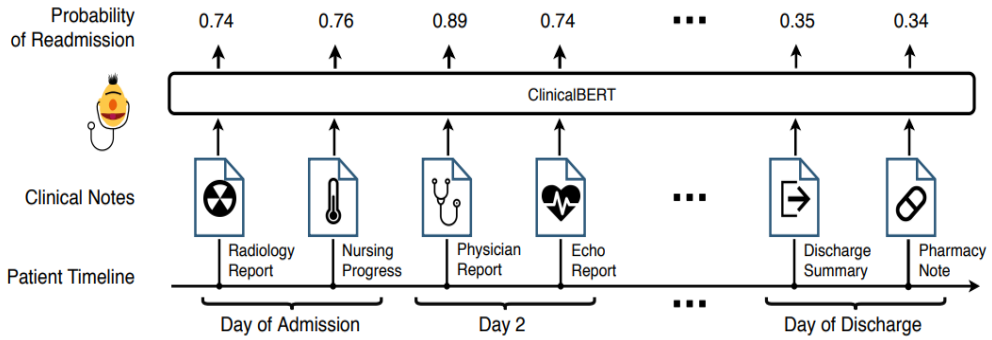
Figure 11. Overview of the pre-training and fine-tuning of BioBERT



Source: Lee et al., 2020, p.1235.

BioBERT performs well in tasks such as NER and question answering compared to traditional methods. ClinicalBERT, which is intended for unstructured clinical text analysis, shows high performance while predicting the readmission of patients to the hospital (Huang, Altosaar, et al., 2020, p. 2), as depicted in Figure 12. In Table 2, ClinicalBERT has been compared with other models in terms of its ability to predict 30-day hospital readmissions to demonstrate its efficiency in the enhancement of clinical decisions.

Figure 12. ClinicalBERT for readmission prediction from clinical notes



Source: Huang et al., 2020, p. 2.

Table 2. ClinicalBERT 30-day readmission prediction accuracy

Model	AUROC	AUPRC	RP80
ClinicalBERT	0.714 ± 0.018	0.701 ± 0.021	0.242 ± 0.111
Bag-of-words	0.684 ± 0.025	0.674 ± 0.027	0.217 ± 0.119
BI-LSTM	0.694 ± 0.025	0.686 ± 0.029	0.223 ± 0.103
BERT	0.692 ± 0.019	0.678 ± 0.016	0.172 ± 0.101

Source: Huang et al., 2020, p. 5.

2.5.2 Evaluating BERT-Based Models for Medical Text Classification

The expansion of biological and clinical data has led to the necessity of improving text classification in health care organizations. The linguistic model BERT has shown highly encouraging performance in different NLP tasks, including text categorization. This review will explore the developments made in the past year in the application of BERT-based architectures for the classification of medical text. Several studies have investigated the efficacy of BERT-based models in medical text categorization, with an emphasis on domain-specific pretraining and architectural modifications that enhance performance.

a) Domain-Specific Pretraining:

Domain-specific pretraining, as evident from CovBERT, shows that fine-tuning BERT for biomedical and scientific document subclassification is quite efficient. The model CovBERT that was trained on the COVID-19 scientific papers dataset was tested, and it scored 94% accuracy after the 4th epoch. This domain-specific pretraining enhances the model's performance and, at the same time, makes the adaptation process more efficient and faster, thereby reducing the learning cost, model size, and training time. The approach shows how it is possible to build deep learning models for specialized tasks using domain-specific information and overcome such problems as a limited vocabulary and small texts (Khadhraoui et al., 2022, p. 7,17).

b) Fine-Tuning with Limited Data:

The study conducted by X. Li et al., (2022, p. 2) aims at the feasibility of fine-tuning BERT based on minimal health NLP data for disease diagnosis. Although the health sector is a challenging environment for learning from limited labeled data, BERT's ability to perform contextual word representation is promising. The study demonstrates that BERT can attain better results even with fewer labeled documents per class through learning curve analysis on disease categorization tasks with two Chinese patient corpora. It also highlights the possibility of BERT surpassing other models in a short period of time as more labeled data is obtained and the model's ability to extract detailed information from each training instance. The findings of this work are highly valuable for understanding the generalization performance of BERT in health NLP issues with limited training samples.

c) Specialized BERT Models:

Med-BERT showcases an understated adaptation of transformer architecture to structured EHR. Med-BERT, unlike traditional BERT's 1-D word sequence input, combines multilayer and multi-relational EHR data via three separate embeddings: code, serialization, and visit embeddings. It should be noted that this specific solution addresses the problem of how one can effectively encode EHR structures. Such pretraining tasks as the Masked Language Model and Prolonged Length of Stay prediction show the bidirectional features of Med-BERT and give contextualized embeddings, which are crucial for the downstream disease prediction tasks. The evaluation on different cohorts shows that Med-BERT has a significant performance improvement; thereby affirming that it is an excellent framework for medical text analysis (Rasmy et al., 2021, pp. 4-5).

d) Transfer Learning Approaches

In the study conducted by Qasim et al., (2022, pp. 4-5) , the performance of different transfer learning models, including BERT-base, BERT-large, RoBERTa-base, RoBERTa-large, DistilBERT, ALBERT-base-v2, XLM-RoBERTa-base, Electra-small, and BART-large, was compared in the context of COVID-19 false information, informative tweets, and extremism. These results demonstrate that transfer learning algorithms can be effectively used for assessing medical text data as well as for gaining insights into the classification of online information concerning public health and extremist discourse.

e) Challenges and Future Directions

Despite the favorable results, there are various difficulties and potential for future study in the field of BERT-based models for medical text classification, which are discussed below.

- Data Scarcity:

Most medical text classification problems suffer from a lack of annotated data, which can affect the performance of BERT-based models. Exploring the possibilities of data augmentation approaches and few-shot learning processes might alleviate this issue. For example, Y. Li et al., (2023, p. 6) put forward an enhanced deep learning model using BERT architecture and PSO to enhance the medical text classification on the constrained Hallmarks dataset.

- Interpretability:

These BERT-based models are often referred to as "black boxes," meaning that it is hard to understand how they arrived at the produced results. Thus, designing BERT-based models that can be easily interpreted or integrating various explainability techniques might enhance confidence and usage in the medical domain. This study by (Huang, Garapati, et al., 2020, p. 1) also investigated the challenges of BERT in the classification of long clinical notes, stating the need for more descriptive models.

- Multilingual and Cross-Lingual Capabilities:

Extending the BERT-based models in the case of multiple languages and cross-lingual transfer learning can improve their application in different language environments. Muller et al., (2021, p. 2217) have found that the BERT model, which was pre-trained on the PubMed abstracts and MIMIC-III clinical notes, outperformed other models on the biological language understanding evaluation (BLUE) and possibly has cross-lingual capabilities.

- Multimodal Integration:

Incorporation of multimodal techniques in medical speech recognition is carried out by integrating deep neural networks, recurrent neural networks, and sequence-to-sequence models with feedback from experts. This approach employs a combination of an RNN for classification, acoustic modeling, grammatical checking, and a professional's opinion. More reliability is afforded by character-based tests across the range of sizes of the vocabulary used. Furthermore, when combining BERT-based language models with medical images and structured clinical data, text classification is enhanced because of the variety of data sources. Thus, the holistic approach allows for more precise and subtle analysis of medical texts, which can have transformative implications for research and clinical practice in medicine (Basystiuk & Melnykova, 2022, pp. 4-6).

Future research on medical text categorization using BERT-based models has a lot of potential, especially regarding domain-specific pretraining and special models. Tackling challenges regarding data availability, model interpretation, language translation, and incorporating multiple modalities will be important components for fully using these sophisticated language models in the medical field.

2.6 Explainable AI (XAI) in Healthcare

When it comes to applying AI in health care, XAI is essential to improving the AI's interpretability in diagnosing, prognostic, and decision-making processes. Due to the realistic and well-explained reasoning behind AI's decisions, XAI eliminates the deep learning models' indefiniteness and, therefore, gains clinicians' trust. Transparency offers the capability to translate AI models' internal structure to users, while explainability transforms AI judgments into human-interpretable forms, thus maintaining trust and accountability in medical applications (Al Kuwaiti et al., 2023, pp. 4-6).These components enable the incorporation of AI into clinicians' work, improve security for individuals, and promote ethical standards in AI-enabled medicine (Bharati et al., 2024, p. 1430).

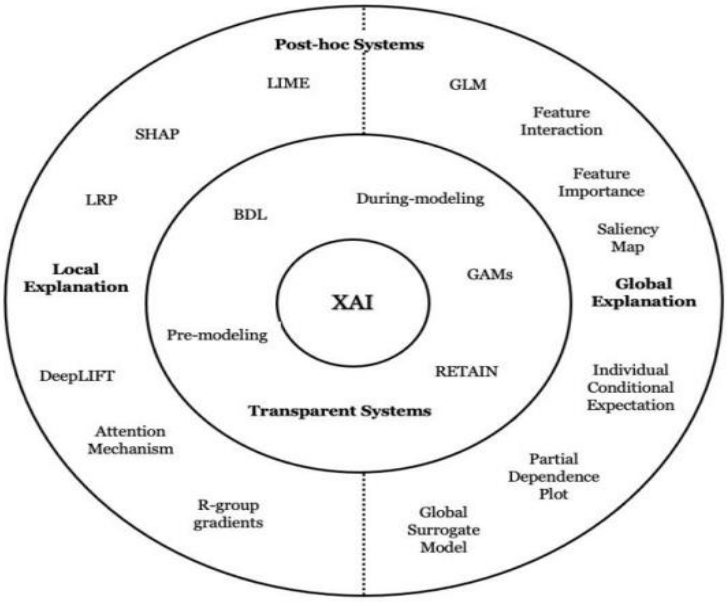
The significance of transparent AI models in healthcare is that they provide greater trust and accountability through clear decision-making procedures and governance that addresses ethical concerns (Al Kuwaiti et al., 2023, p. 2). Transparent AI facilitates informed decision-making by providing insights into AI suggestions, allowing clinicians to modify decisions based on patient-specific details (Kanda et al., 2020, p. 2). Additionally, transparency helps identify biases and errors in data and algorithms, fostering robust systems (Nguyen, 2023, p. 1). Therefore, continuous improvement is

attained through the collaboration of medical practitioners, AI specialists, and policymakers to ensure a dynamic loop of feedback and innovations in the implementation of AI in the health care system (Bajwa et al., 2021, p. 190).

2.6.1 XAI Techniques in Healthcare

XAI methods in healthcare improve the understanding of how an AI model reaches its decision and thus increase the likelihood of patients' safety, legal requirements, and ethical standards. Figure 13 illustrates two levels of XAI algorithms: interceptive systems, such as RETAIN and GAM, as well as post hoc systems, including LIME and SHAP (Chakrobartty & El-Gayar, 2021, p. 3). They help clinicians in decision-making, models' interpretability, and visualization tools, thus enabling the integration of AI systems with healthcare professionals.

Figure 13. Explainable methods and techniques categories



Source: Chakrobartty & El-Gayar, 2021, p. 3.

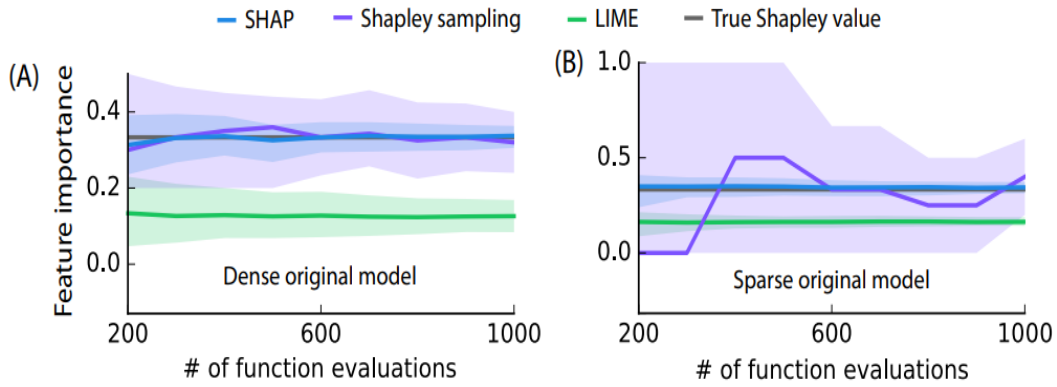
Feature-importance techniques are important for analyzing the AI models' decision-making in health care. These include patterns of patient characteristics or clinical factors that are associated with these outcomes, giving an understanding of how these variables may affect each other and the patient's overall outcomes. In general, feature-importance methodologies' studies seek to refine not only the AI models used in healthcare but also the decision-making process in medicine and the care of patients.

a) SHAP

SHAP values offer an extensive approach to feature importance estimates, and it is a consistent measurement for many of the models. They allocate the differences in the estimated model's prediction to the attributes, explaining the impact of each feature. SHAP values are derived from elaborate techniques such as global and individual interpretable techniques and model-agnostic and

model-specific approximations. Kernel SHAP, for instance, applies weighted linear regression to solve the problem of feature importance with higher accuracy compared to Shapley sampling and LIME.

Figure 14. Comparison of kernel SHAP, shapley sampling values, and LIME additive feature attribution methods



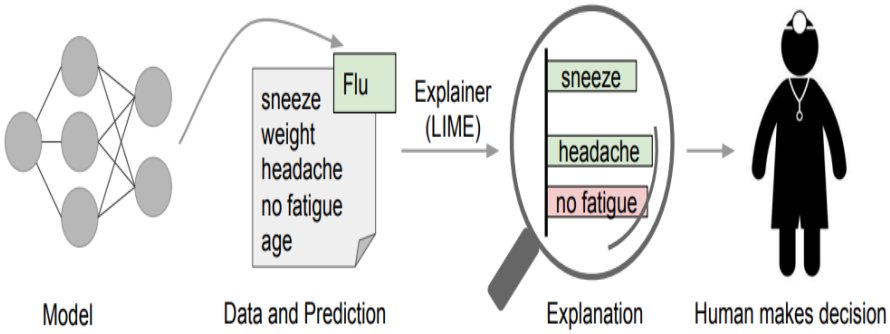
Source: Lundberg & Lee, 2017, p. 8.

Figure 14 compares three additive feature attribution methods: kernel SHAP (where the lasso is replaced with a debiased version of the method), Shapley sampling values, and LIME. When the number of model evaluations rises, the two models' 10th and 90th percentiles of 200 replicate estimations for one feature are displayed. Panel (A) defines a decision tree model where it includes ten inputs, on the other hand, Panel (B) defines a decision tree where it has 100 input attributes, of which only three are selected. Kernel SHAP is also found to be more efficient and accurate than other methods, especially in sparse models. SHAP values also tend to be closer to human-like reasoning, thus aiding in a more credible model explanation than LIME and DeepLIFT. Combining SHAP values with other techniques such as DeepLIFT improves interpretability in complex deep learning models such as the MNIST digit classification (Lundberg & Lee, 2017, p. 8).

b) LIME

LIME is a technique in the broader field of XAI to offer reasonable explanations for certain predictions of a black box model through locality. It explains examples in a format that can be easily understood by the user, such as binary vectors on text or image data inputs to the model, while enabling the user to understand the model's decision process despite its complexity. LIME removes the trade-off between accuracy and interpretability by narrowing the distance between the opaque model's predictions and an explainable surrogate model in the vicinity of the instance being explained. That is done through sampling and sparse linear models; thus, it can be used for text and image classification, where the representation can be, for example, the bag-of-words or super-pixels. With an emphasis on exact predictions, LIME improves the interpretability of black box models and, therefore, the confidence in them in different fields (Ribeiro et al., 2016, p. 1136). Figure 15 illustrates the process of explaining flu predictions.

Figure 15. Explaining predictions with LIME: flu example



Source: Ribeiro et al., 2016, p. 1136.

c) Word Importance Scores

Importance scores of the words are one of the most crucial XAI techniques that can help in the identification of the impact of the specific words in the text. These scores include techniques like TF-IDF, or word embeddings, that take into consideration elements such as the frequency and relevance of the context when they rank the importance of each word. Word importance scores as an example of the non-traditional attention mechanisms are more complex than the basic attention values since they disassemble the word significance into such measures as bias, reading level, or verbosity (Asudani et al., 2023, pp. 10352-10354). his enhances the understanding of language models since the analyst can assess the extent to which each word impacts the model's outcome. Such findings have implications in many contexts, such as text summarization, document classification, and adversarial attacks where the identification of significant words is important. Such techniques as, for example, CRank have been implemented to estimate these scores, which proves their applicability in tasks where textual input is analyzed in detail (X. Chen & Liu, 2021, p. 5). In a nutshell, word importance scores enable the understanding of the relevance of a certain word, which in turn contributes to the enhancement of the language models' and decisions' interpretability.

d) Topic Modeling

Techniques in topic modeling, including Latent Dirichlet Allocation and its variations, ProdLDA, and NeuralLDA, are useful in modeling texts since they represent documents and topics as distributions. These methods give more easily interpretable results in terms of the themes that are present in texts. Integrating topic modeling in XAI frameworks increases understanding of how specific outcomes in text categorization are arrived at, increasing the models' transparency. Solutions to those problems include the Cross-Lingual Contextualized Topic Model and the Embedded Topic Model, which expand the XAI spectrum of applications (Rijcken et al., 2022, pp. 3-6).

e) Attention Mechanisms

In the healthcare context, attention mechanisms in XAI systems improve interpretability and visualization through the global, local, and self-attention processes to build up the context and look for the

relations between inputs. Self-attention mechanisms are particularly helpful in forecasting clinical events and evaluating their severity, particularly within the ICU environment. RETAIN and its derived versions, like RetainVis, retain interpretability as they help to consider the effect of each visit and variable, which is helpful for incorporating sequential data and maintaining therapeutic significance. Attention-based XAI techniques, including Grad-CAM, are used in medical imaging for tasks such as appendicitis detection and hypoglycemia diagnosis to help the physician pinpoint important regions. However, issues such as information overload and warning fatigue underline the necessity of maintaining the right proportion in delivering information to healthcare providers (Bharati et al., 2024, p. 1435).

f) Model Visualization Techniques

The techniques of model visualization aim at providing users with a visual representation of the inner structure of an AI model so that they can understand the model's behavior and decision-making. In healthcare, model visualization techniques have been employed in the analysis of the model's output and the search for patterns or anomalies in the input data. Two major methods for visualization are mentioned below:

- Grad-CAM: Grad-CAM (Gradient-weighted Class Activation Mapping) improves the comprehensibility of CNN by drawing attention to the areas in the input images that are significant for the decision-making process. They adopt gradients entering the final convolutional layer to calculate neuron importance for a given class to produce class discriminative localization maps. Compared to previous approaches, Grad-CAM offers high-resolution visualizations without modification of the architecture and is suitable for different applications such as image description and answering questions about images. Grad-CAM, when coupled with guided backpropagation, provides high-resolution visualizations and helps in understanding why a particular decision was made by the network, thus building faith in the automated systems (Selvaraju et al., 2020, pp. 339-340).
- TCAV: TCAV (Testing with Concept Activation Vectors) is one of the XAI techniques that may be used to track the neural network model sensitivity to the concepts specified by the user. It operates by determining Concept Activation Vectors, which encode these concepts within the model's activations, and then using TCAV scores to estimate these concepts' influence on predictions. This method allows assessing the model's behavior without retraining, does not allow training in irrelevant concepts, and makes it possible to compare the importance of different concepts (Kim et al., 2018, pp. 2-4). Overall, TCAV provides valuable insights into model behavior and ensures the reliability of explanations

g) Word Importance Visualization

Word importance visualization is one of the prominent techniques in XAI that aims at visualizing text-data comprehension. Its main purpose is to identify the most significant words or tokens that are important for a model's prediction, thus improving the model's interpretability. A few methods have

been proposed to quantify word importance, which are followed by saliency maps obtained from the XAI algorithms (Adebayo et al., 2018, p. 4). Transformer-based models can greatly benefit from the interactive self-attention visualizations as well as the template-based systems given controlled explanations. However, each of the strategies presented above has its strengths and weaknesses. For example, to overcome the problem of spurious correlations or the ability to obtain similar results, continuous developments, including the creation of hybrids, try to overcome these drawbacks. Nonetheless, word importance visualization still remains a major approach in XAI for text categorization, which tries to explain the model's decision-making by highlighting the importance of words (Clement et al., 2023, p. 86).

h) Word Embedding Visualization

The t-SNE and PCA techniques used in XAI are important for mapping and revealing the semantic similarity of NLP models. These techniques map high-dimensional word vectors to lower-dimensional space, which helps in easy comprehension of the semantic relations and meanings of the words in the given text (Heimerl & Gleicher, 2018, p. 254). Such methods as the combination of topic modeling with word embeddings improve the interpretability of the NLP models. However, visualization should complement quantitative evaluation methods, not substitute them. The selection of the dimension reduction and data selection choices has a direct impact on the visualization's interpretability and usefulness; thus, it is important to complement these techniques with other assessment methods to enhance the model’s explainability and decision-making support for AI-based systems (Oubenali et al., 2022, pp. 3-4).

i) Challenges and Future Directions

The difficulties and future perspectives of XAI approaches for medical applications are essential for enhancing the transparency and interpretability of AI decision-making. Here are the details of these factors:

- Bridging the Gap between AI Developers and Clinical Users: Effective interaction among AI developers and physicians is essential to tailoring XAI models to the specific needs of healthcare. This requires developing ways to encourage interaction and comprehension among technical experts and domain specialists
- Addressing Legal and Ethical Concerns: XAI techniques must contain protocols for addressing legal and ethical issues like data privacy, algorithmic bias, and responsibility. Future research must focus on developing ways that enhance fairness, accountability, and openness in healthcare AI systems (Jung et al., 2023, p. 7). Addressing these challenges is critical to the effective deployment of XAI in clinical practice.
- Evaluating Explanation Effectiveness: It is vital to create reliable methods for evaluating the efficacy of XAI model explanations. These methods should consider the perspectives of both

physicians and patients, ensuring that explanations are not only understandable but also valuable in clinical decision-making.

- Integrating XAI into Clinical Workflows: To implement XAI models in the present clinical practice, approaches should meet the concerns of the users, organizational limitations, and technical specifications. Subsequent research should aim at developing design methodologies that support the implementation of XAI in clinical environments.
- Advancing XAI Techniques: More sophisticated XAI methods must be investigated and improved to meet the complexity of medical data and provide detailed explanations. This includes exploring how best to deal with different data types, time series data, and data from multiple sources, and improving the interpretability of complex models such as deep learning models (Frasca et al., 2024, p. 3).

Overall, addressing these difficulties and future developments in XAI techniques is crucial to realizing XAI's full promise in medical applications, which includes improving patient outcomes and clinical decision-making.

2.6.2 Effectiveness of XAI in Medical Text Classification

As mentioned in the above studies, XAI is now considered a significant field in the medical domain, where the explainability of AI-based decision-making systems is crucial. In the context of medical transcription categorization, XAI techniques are essential for doctors and other medical personnel to understand the rationale behind the AI model's decisions. This review aims at assessing the performance of XAI by comparing various methods and also assessing the advantages and shortcomings of the XAI approaches in the medical transcription categorization problem area.

a) Comparison of XAI Techniques for Medical Transcription Classification:

i. Model-Agnostic XAI Methods:

The following is a comparison of three model-agnostic XAI methods: LIME, SHAP, and LRP in the setting of medical transcription categorization (Zhou et al., 2023, pp. 3-5).

- LIME: Explanations for ML models at the local level are produced by selecting a data point, using that data point to create new instances by making small changes to it, training a linear model on the new instances, and minimizing an objective function to estimate the behavior of the complex model at the local level. The object function is demonstrated below.

$$argmin_g \sum_{i=i}^{n} \omega_i \, L\big(f(x_i), g(x_i)\big) + \Omega(g)$$

where $f(x_i)$ is the complex model prediction on an instance x_i, g(x_i) is the predicted outcome of the linear model on instance xi, The loss function L(f(x_i), g(x_i)) measures the difference of the predictions of the complex and linear model, while the weight ω_i is the weight allocated to instance x_i, which is a function of the distance between x_i and x. The weight ω_i increases as the instance x_i becomes closer to x. The regularization term Ω(g) affects the linear model's

complexity. In addition, one should find out how individual cases are included in a calculation that yields predictions and use this information to assess other medical text transcriptions and identify concepts therefrom.

- SHAP: To interpret the outputs of an ML model, Shapley values derived from cooperative game theory are used to distribute each feature's contribution towards the prediction of a single instance. Shows the extent to which each feature contributes to the prediction. For a model f, the Shapley value ϕ_j for the j-th feature is defined as:

$$\phi_j(x) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f_{S \cup \{j\}}(x) - f_s(x)]$$

Where N = 1, 2,..., n is the set of all features, and |S| signifies the cardinality of the set S. |N| denotes the set's cardinality. $f_s \cup j(x)$ is the model's prediction based on S features and the j-th feature. f_s(x) is the model's predicted value when just the features in S are taken into consideration.

Moreover, it is suitable for comparing the value of features in medical text transcription data that are useful in identifying notable factors that affect the classification outcome.

- LRP: Layer-wise Relevance Propagation provides an explanation for the deep neural networks' predictions by assigning relevance scores to the input features. Details regarding the value of features or neurons can be obtained when relevance scores are propagated backward through the layers of the network using a specific formula. For instance, for the last layer's k-th feature map with width m and height n, the average value is the sum of all activation values $f_k(x, y)$ in the k-th feature map is defined as:

$$R_j = \sum_k \frac{a_j \omega_{jk}}{\sum_j a_j \omega_{jk}} R_k$$

Further, it is implemented in the text classification of medical transcription applications for deep learning models, which gives the specifics of which input factors affect the network's decision-making mechanism most.

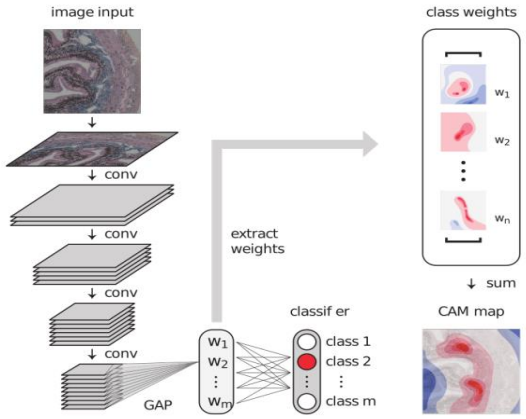
ii. Model-Specific XAI Methods:

When it comes to model-specific methods for XAI in medical transcription classification, two techniques stand out: class activation maps and attention scores are used. Figure 16 depicts how CAM leads to better interpretability through the generation of heat maps that indicate the part of the image to be used by a convolutional neural network in reaching a given decision. By calculating the weighted feature mappings in the last convolutional layer using the following equation: For example, for the k-th feature map of the final layer with a width m and a height n, the average value is the sum of all activation values f_k(x, y) in the k-th feature map.

$$\omega_k = \frac{1}{m\,n}\sum_{x}^{m}\sum_{y}^{n} f_k(x,y)$$

CAM effectively illustrates where, in the input image, the model focuses to make its decision. This approach allows healthcare professionals to determine which features or patterns in the medical images matter in the classification task and helps in diagnosis as well as the preparation of the best therapy strategies (Zhou et al., 2023, p. 5).

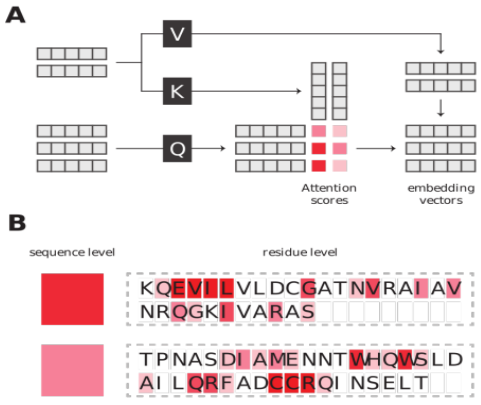
Figure 16. Procedure of CAM



Source: Zhou et al., 2023, p. 5.

Attention scores, shown in Figure 17, offer a new form of interpretability. Attention-based models are more informative in terms of the model's decision-making process because it is possible to quantify the importance of input components to a single element of output. The attention scores can be visualized by the medical practitioners to identify which input tokens (i.e., words or phrases in medical transcripts) were more influential in the classification. This allows not only for the understanding of the global prediction but also for the detailed clinical phrases or contexts that shaped the model's decision, enhancing accountability and confidence for the classification results for medical transcriptions (Zhou et al., 2023, p. 5).

Figure 17. Example of attention scores



Source: Zhou et al., 2023, p. 5.

iii. Ensemble or Multimodal XAI Approaches:

Ensemble XAI techniques play a significant role in improving the accuracy and interpretability of the medical text classification models that are fundamental to health diagnosis and treatment. This research combines several XAI approaches, such as GradCAM, LIME, and saliency maps, to work on and enhance medical text data. GradCAM provides coarse localization maps, LIME provides locally interpretable models, and saliency maps evaluate pixel-level influence, altogether solving the problem with individual methods. In this way, the study will be able to provide a holistic view of the patterns of medical data, which will in turn help to improve the identification and classification of diseases (Chaddad et al., 2023, p. 11). Here, the study goal is to determine the best XAI algorithms to improve diagnostic precision and clinical decision-making.

b) Strengths and Limitations of XAI in Medical Transcription Classification:

- Improved Transparency: XAI approaches can enhance the decision-making process of the categorization models used in medical transcription, allowing medical personnel to understand the logic behind the system's output. This can raise confidence and facilitate the implementation of this technology in clinical processes.
- Identification of Influential Features: SHAP and CAM of XAI approaches may identify those that are most influencing features of language or medical concepts in input transcript data, which could be useful for further development of models and improvement of the refinement process (H. Zhang & Ogasawara, 2023, p. 8).
- Multimodal Explanations: Combining different XAI approaches can offer a broader and deeper understanding of the model's decision-making process while also taking into consideration the drawbacks of each method (Park et al., 2018, p. 8781)

However, the use of XAI in medical transcription classification has significant limitations:

- Complexity of Medical Language: Medical terminology is unique, with expert terminology and intricate word formations and relations that, in general, do not allow one to fully understand the approaches to XAI. Creating the XAI algorithms that may potentially be able to capture these language definitions is still in its research phase (Zhou et al., 2023, p. 13).
- Scalability and Computational Efficiency: Certain XAI techniques, including SHAP, might be computationally intensive, mainly when analyzing extensive medical transcription datasets. In other words, achieving an optimal solution to the trade-off between the quality of explanations and computation time is an ongoing process (Holzinger et al., 2022, pp. 16-17).
- Validation and Clinical Integration: The categorization of medical transcription in the healthcare process must be considered clinically relevant, and the XAI-based systems should be trusted. However, validation of these technologies in real-world situations means that integration of these technologies into practice needs more elaborate testing and user analysis.

Overall, the results of the search show an increasing focus on and use of several types of XAI in the context of medical text classification. Explaining medical AI decisions with model-agnostic and model-specific XAI methodologies, as well as the creation of multimodal explanation XAI approaches, shows the number of strategies being explored to enhance the interpretability and trustworthiness of medical AI systems.

2.7 Integration of BERT and XAI Techniques

The integration of BERT and XAI enhances the accuracy, interpretability, and reliability of medical text transcription classification models. The improvement of contextual and semantic comprehension of medical texts in BERT improves the identification of medical terminology, symptoms, and diagnoses, hence improving categorization accuracy and context sensitivity. This, in turn, increases the quality of patient treatment and clinical outcomes (Binder et al., 2022, p. 2124). The integration of XAI methods, including saliency maps and feature attribution to BERT, enhances the model's interpretability and trustworthiness, which is vital for healthcare professionals to accept AI-based solutions (Mavrepis et al., 2024, p. 5). This makes it easier to provide explanations for the model's predictions and enhance the model based on the feedback received.

Furthermore, the combination of BERT with XAI approaches allows for the detection of errors and discrepancies in medical texts, thus increasing the credibility of clinical records. This enhances transparency to help in compliance with rules and ethical codes of conduct, eradicating bias and error. Moreover, XAI methods help to continue improving the model since its weaknesses and biases are explained, thus improving the quality of analyzing medical texts (Szczepański et al., 2021, p. 2). This is where BERT and XAI can complement each other and lead to profound improvements in the field of medical text categorization, with positive impacts on patients' outcomes and clinical decisions.

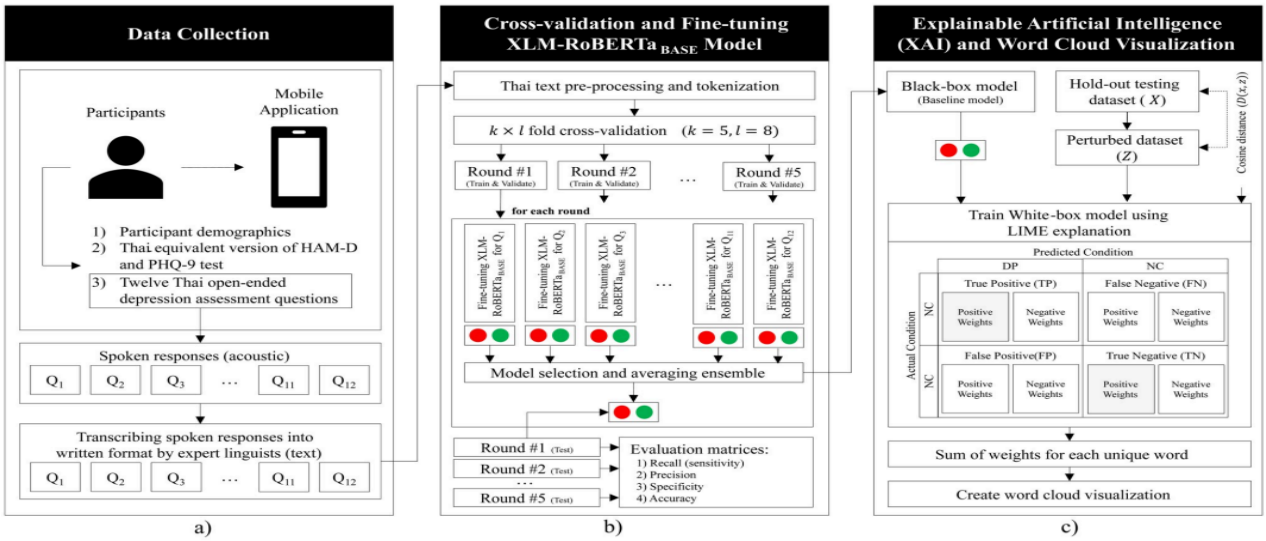
2.7.1 Studies and Applications of BERT and XAI Integration in Medical Text Transcription Classification

The integration of BERT and XAI methods has been considered for medical text transcription categorization. BERT, a pre-trained language model, is highly effective in capturing the context of medical text with respect to positive and negative sentiment and symptoms. It can address bidirectional dependencies in text since BERT is designed to process text bidirectionally; other techniques such as LIME can also help in the interpretation of BERT, thus making the model's outputs more explainable and reliable. Figure 18, "A flow diagram: Data Processing, Model Training, and LIME Explanation," presents the entire flow from the data collection to model tuning and the incorporation of BERT with XAI, thus explaining the systematic nature of this area of research.

In medical applications, it is found that the integration of BERT and XAI is beneficial, especially in the diagnosis of mental health. The implementation of BERT with small sets of transcripts of textual responses has also been effective in detecting depression and acknowledging the experience. The reviewed XAI methods, such as LIME, improve the transparency of model predictions through

explanations that physicians can understand to determine the factors that led to diagnoses. Furthermore, the studies related to the gender differences in depression prove that gender-specific methods are significant in medical text classification; therefore, the application of BERT and XAI is advantageous for both accurate diagnosis and gender-specific insights (Munthuli et al., 2023, pp. 3-4).

Figure 18. Flow diagram: data processing, model training, and LIME explanation



Source: Munthuli et al., 2023, p. 5.

In another study, enhancing the BERT models combined with the XAI techniques advanced medical text classification. RoBERTa, ClinicalBERT, BioBERT, and other models based on BERT are widely used in the present NLP tasks because of the contextual understanding of the models. Interpretable methods such as integrated gradients assist in explaining BERT's predictions by providing importance scores to words in the text. This combination improves the interpretability of the models by making them easier to understand and increases the confidence that is placed in the model predictions. Also, the application of fine-tuned BERT models in the medical imaging protocol assignment tasks was more accurate compared to traditional machine learning approaches, while the XAI methods helped in understanding the in- and out-coming parts of the model, revealing the increased classification accuracy and model interpretability (Talebi et al., 2024, p. 3).

2.7.2 Practical Improvements in Integration of BERT and XAI Techniques into Clinical Practice and Decision Making

The use of BERT, along with XAI techniques, to integrate them into the clinical workflow and decision-making process can positively impact the health care industry. The combination of these technologies provides prospective options for healthcare developments since it gives clinicians and healthcare professionals interpretable and believable models. Below is a description of practical improvements in clinical settings.

- Improved Clinical Decision-Making: By applying XAI methods to explain BERT's decisions, clinicians can better comprehend the AI's decisions and make more competent decisions about the results of treatment for their patients.
- Enhanced Patient Safety and Outcomes: Utilizing XAI to identify bias and discrepancies in AI models enables not only the dependability of BERT in healthcare but also actively contributes to patient safety by minimizing the possibility of serving biased recommendations, ultimately encouraging users' trust in the system (Moazemi et al., 2023, p. 15).
- Increased Trust and Adoption of AI in Healthcare: Increasing the interpretability and explaining capabilities of BERT with XAI allows for its higher acceptance among healthcare workers, creating the context for the further adoption of AI solutions in clinical practice and, therefore, the improvement of patient care.
- Streamlined Clinical Workflows: Utilization of AI-driven decision support technologies including BERT in clinical practices automates these repetitive tasks hence giving clinicians more time to handle complex cases and increase on the efficiency of the healthcare systems (Elhaddad & Hamam, n.d., p. 4).
- Personalized and Precision Medicine: AI enables the processing of big data to recommend treatment plans and the prognosis of the patient, improving the concept of personalized medicine (Johnson et al., 2021, p. 87).

2.7.3 Challenges and Future Directions: Integration of BERT and XAI Techniques

While the utilization of BERT and XAI techniques in clinical settings shows significant promise, there are numerous obstacles that must be addressed:

- Interpretability vs. Performance: It is important for the model based on BERT to have high predictability, while the application of the XAI approach should offer satisfactory interpretability. Researchers must create techniques that would maintain high accuracy and, at the same time, make an AI system more transparent so that physicians could trust the model and understand why it made a certain decision.
- Clinical Validation and Acceptance: These technologies require extensive clinical testing and acceptance by health care professionals due to their broad application. To ensure that AI systems meet clinical needs and deliver useful information and solutions for resolving real-world healthcare concerns, intensive collaboration with technical specialists and clinicians is required.
- Data Privacy and Ethical Considerations: Privacy and ethical issues concerning the use of sensitive patient information must also be solved. Evaluating the risks and practicing proper data anonymization techniques, as well as following the principles of responsible AI development, are ways to minimize the risks, preserve the confidentiality of patients' data, gain their trust, and meet ethical standards (Murdoch, 2021, pp. 1-2).

- Scalability and Integration into Clinical Workflows: The adoption of BERT and XAI approaches in clinical practice raises certain issues about the feasibility and universality of improving present clinical operations. AI is only useful in healthcare if it is applied to systems that can instantly interact with existing systems and deliver the clinical solutions required across all disciplines. The successful implementation of AI technology in healthcare requires the development of flexible systems (Barragán-Montero et al., 2022, pp. 6-8).

The implementation of BERT and XAI for medical care means that some important issues must be considered, namely, the data privacy issue, the issue of model validation, and the issue of compliance with existing regulations. The successful application of AI systems is based on the coordination of healthcare practitioners, AI developers, and governing bodies. The integration of both methods can contribute to improving medical text transcription and classification since their models can be more accurate, interpretable, and reliable. Thus, by making BERT's "black box" decision-making more understandable and traceable, this approach may benefit patient safety, clinicians, and other stakeholders in the healthcare system who aim to incorporate AI into clinical practice.

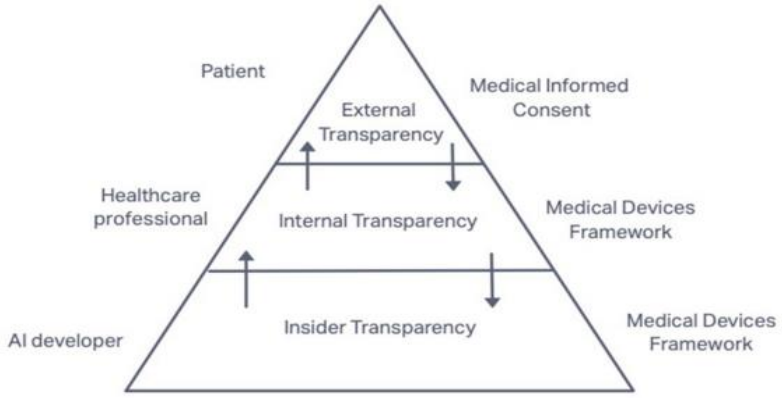
2.8 Ethical and Regulatory Considerations

In the dynamic environment of the healthcare system, the use of AI brings up several ethical and legal issues. With the integration of AI in various aspects of healthcare service delivery, issues of transparency, fairness, and accountability gain more prominence. Ethical issues are related to the design, deployment, and monitoring of AI-based health care systems and shape legal requirements and business standards. Managing these aspects, starting with patient rights and ending with algorithmic bias, is vital for gaining trust, achieving equality, and enhancing patient safety. This part focuses on the interaction between ethical guidelines and legal requirements and provides insights into the basic principles and approaches that define the proper application of AI in medicine.

2.8.1 Ethical Considerations Related to Transparency, Fairness, and Accountability in AI-Driven Healthcare Systems

It is important to note that transparency is a vital component in the AI-supported healthcare system because it determines the level of accountability, safety, and quality. It empowers stakeholders to comprehend how AI makes decisions, stay in charge, and ensure that algorithms are accurate. This includes interpretability, explainability, auditability, and documentation measures that support accountability and risk management, as shown in Figure 19. Transparency in AI has to be properly adjusted for the medical field and incorporated into the existing systems of accountability for efficiency (Kiseleva et al., 2022, pp. 9-10).

Figure 19. Multilayered system of AI's transparency in healthcare



Source: Kiseleva et al., 2022, p. 10.

In AI-powered medicine, fairness means that there is a balance in the distribution of resources, opportunities, and results for patients of different demographics. This involves data bias, which results from the use of inadequate or skewed data, and algorithmic bias, which originates from the structure and training of AI systems. To increase fairness, it is necessary to employ various data sets, perform checks on AI algorithms from time to time, increase awareness among stakeholders regarding AI biases, and improve data privacy. There is a need for a well-articulated set of rules and guidelines that would determine the responsibilities of doctors, AI developers, companies, governments, and patients in the context of AI-assisted health care. Physicians need to verify the diagnosis produced by AI, while AI developers must ensure the model's reliability and non-bias. Healthcare organizations should accept and evaluate AI solutions, and governments should set rules for AI implementation. All stakeholders must devise strategies for eliminating biases, increasing trust, and improving patient outcomes (Ueda et al., 2024, pp. 4-8).

2.8.2 Regulatory Frameworks and Guidelines Governing the Use of AI in Healthcare, Particularly Concerning Transparency and Accountability

Policies and standards are crucial to enhancing the safe and sustainable use of AI in medical environments. The General Data Protection Regulation (GDPR) of the European Union offers strong protection of personal data and focuses on the transparency of the decision-making mechanisms in the case of automated decisions. The proposed AI Act (AIA) goes further by adding specific restrictions pertaining to AI technologies, which take into account the differences in risk levels and attempt to strike a balance between innovation and safety (Schöffer, 2023, p. 8). In the U.S., the FDA's guidelines for AI/ML-based medical devices highlight transparency, validation, and monitoring in compliance with GDPR principles to ensure the safety and efficacy of AI systems and to build trust among healthcare workers and patients (Health, 2024, p. 3). The American Medical Informatics Association offers principles of fairness, accountability, transparency, and ethics in AI-enabled healthcare to help practitioners and policymakers deal with ethical issues while focusing on the patient's best interests (Singhal et al., 2024, p. 2).

However, there are some issues that still persist. For instance, unethical individuals often exploit the time lag between technological advancements and the implementation of legal frameworks. There are issues of resource constraints, especially among small-scale medical organizations, and there is a need to foster cooperation among countries to standardize the laws and improve the health of the people in the world. Transparency and accountability are the key principles for AI governance, along with ethical guidelines and cooperation with other countries.

2.9 Research Gap

There is tremendous progress in the use of BERT-based models in the classification of medical text transcription, but there is still a significant research gap. Although some models like BioBERT and Clinical-BERT have been developed, there is no study that compares their effectiveness across different medical specialties. This difference indicates that these models should be compared to identify which of them is more appropriate for each specialization, given that each field has its own terminology and structure. Moreover, current studies often focus on the model's performance in terms of accuracy, while the questions of interpretability and explainability, which are essential for clinical application, remain insufficiently explored. XAI methods like SHAP, LIME, and attention models are still more or less unknown, and their efficiency for explaining intricate medical predictions or their usage is not widely researched. Also, the researchers need to explain how these models deal with the specialty-oriented terms and how the differences meet the requirements to ensure both accuracy and adoption by medical professionals.

In addition, no studies have been conducted in which both the traditional machine learning models and the transformer-based models have been tested and compared systematically on standardized datasets in terms of several performance indices. It is even more crucial to underline the need for a robust methodological approach to assess the efficacy and readability of such models in an impartial manner. Clinical notes lack structure, the heterogeneity in terminology used in clinical notes, and the general shortage of annotated datasets constitute important challenges that the present study does not fully solve. The future work should focus on enhancing the preprocessing techniques for the specific domain, applying more sophisticated NLP techniques to parse the medical terminology, and studying new data augmentation strategies to augment the training data for the models. Another way is to involve clinical professionals to ensure that the developed models reflect the real conditions and are not distorted by a particular researcher's vision. Eliminating these drawbacks will lead to more precise, transparent, and reliable AI medical text transcription categorization that enhances the clinical decision-making process and patients' care. They could result in better customized treatments that could enhance general treatments for patients.

2.10 Summary

The literature review discusses the significance of medical text transcription in health care, emphasizing its use in communication, maintaining records, and clinical decision-making. This approach emphasizes the difficulties in identifying medical texts, given the enormous scope of the healthcare sector and numerous types of documents. The review also describes the development of medical transcription from manual procedures to modern AI-based methods and the future development of AI, ML, and NLP. Other ML algorithms such as Naïve Bayes, Support Vector Machines, and deep learning such as BERT can also be used in automating the classification of medical text, but their major drawback is that most of them are black box models.

XAI helps overcome this problem due to its ability to enhance the levels of transparency and trust in the AI models that are crucial for applying the models in medicine. Clinical text analysis has greatly benefited from BERT and other similar models like BioBERT and Clinical BERT due to their proficiency in understanding highly technical medical language. However, XAI methods, including SHAP, are still not widely used, which makes it necessary to perform multiple investigations to define the perspective of the introduction of this concept into practice. The literature review study shows that there are no systematic studies on the performance of different BERT-based models in different medical specialties, emphasizing the need to use accurate methodological approaches to compare and explain the models' performance. The unstructured nature of clinical notes, together with the lack of annotated datasets, pose substantial challenges that require the need for novel preprocessing techniques and data augmentation specific to clinical notes. Ongoing interaction with clinical personnel is important to make sure that the models are realistic as well as useful for enhancing the quality of clinical decisions and the outcomes for patients. Improving these aspects would enhance the reliability, legitimacy, and effectiveness of such AI-based solutions for the categorization of medical text transcription to enhance patients’ health and the advancement of precision medicine.

Chapter 3: Research Methodology

3.1 Introduction

This study investigates the ability of the BERT pre-trained language model to classify medical text transcriptions from diverse disciplines, solving substantial gaps in the previous literature. Specifically, the study underlines the importance of conducting a systematic performance comparison of BERT variations and incorporating XAI methodologies to improve interpretability and trust in clinical contexts. Using datasets from mtsamples.com, the study uses a combination of classic machine learning techniques and advanced transformer models like BERT, BioBERT, ClinicalBERT, and RoBERTa. To ensure data quality and consistency, comprehensive data preprocessing, EDA, and thorough model evaluation are used. The goal of adopting XAI approaches is to improve the accuracy, transparency, and clinical application of NLP models in healthcare. Finally, this study is aimed at contributing to more reliable AI-driven medical transcription categorization by providing key insights and dealing with ethical issues in the discipline of medical AI.

3.2 Research Approach

Studies being researched in this work seek to evaluate the performance of BERT-based models in classifying medical text transcriptions based on the literature review findings. The thesis work is initiated by defining goals and measures that will help to understand the further effective work of different BERT versions to consider the intricacies of the medical language and its hierarchical structure with reference to various medical fields. The strategy laid down also focuses on quantifiable objectives, thus ensuring the applicability and impact of the study on the advancement of medical AI. The investigation remains rigorous in data collection, exploratory data analysis, and data preprocessing to ensure quality and consistency. The proposed research study is to identify the best performing techniques in medical text categorization by comparing classical machine learning algorithms like Random Forest, XGBoost, Logistic Regression, and SVM with state-of-the-art transformer models including BERT, BioBERT, ClinicalBERT, and RoBERTa. XAI techniques enhance the understanding of a model's prediction results, which is crucial for clinical practice. The intent of the research is to improve model trustworthiness and provide actionable insights for creating NLP applications in healthcare through rigorous misclassification analysis and validation. Thus, this regulated and systematic process helps in making the findings more dependable and beneficial to the field.

The research design is based on the theoretical framework established in the literature review section, where gaps are identified and the basis for an empirical study is created. This ensures that the outcomes of the study are both methodologically correct and relevant to practice since it defines study goals that are linked to the actual requirements for medical transcription work. So, the work ensures the hypotheses about the effectiveness of using BERT-based models in medical text classification through the empirical methods of data collection, rigorous analysis, and validation are well grounded. This methodical approach also enhances the credibility of the results and contributes to

the development of NLP methods for assisting medical practices. Lastly, this research strategy facilitates an understanding of the relationship between theoretical conceptions and empirical evidence to close the gap between theoretical and practical AI in medical settings and enhance the clinical transcription systems' transparency, reliability, and efficiency.

3.3 Research Design

Figure 20. Design of thesis research methodology

Source: Own representation.

The design of the research depicted in Figure 20 carefully examines the BERT pre-trained language model in categorizing medical text transcription. The first step is data collection, where the number of sources is taken from the medical transcription field to gather large amounts of data, and the second step is exploratory data analysis, where the dataset is analyzed to understand the distribution of the data and the patterns of the data. This lays the foundation for effective data pre-processing decisions, which include text normalization and tokenization to produce quality data for model training. The approach outlines the choice of BERT model versions such as BioBERT, Clinical BERT, and RoBERTa and adjusts the hyperparameters to match the specificities of medical transcription tasks. This kind of integration helps the model be fit for the medical language's complexity, improving its accuracy and applicability across various medical specialties.

The models are compared to conventional machine learning algorithms and deep learning models, and strict assessment criteria are applied to ensure that their performance is thoroughly tested. The application of XAI techniques is particularly relevant since it improves the model's decision-making process, allowing for the detection of biases or errors. Moreover, by applying XAI in the study, the reliability of the model's predictions is established, and biases or errors are easily identified (Ali et al., 2023, p. 11). This quantitative approach provides reliable and accurate results not only in one dataset but in different settings as well. The study also carefully assessed the credibility and dependability of the models. Thus, laying a solid foundation for future research into explainable and reliable AI applications in the medical field.

3.4 Data Collection Process

Data collection entails the process of acquiring data to accomplish the objectives of a study. It contains defining objectives, selecting approaches, acquiring and developing data for analysis, and adhering to precision, dependability, and integrity. Due to this organized approach, decision-making is well informed, and the insights gleaned are useful (Taherdoost, 2021, p. 12). The goal of the research

study was to improve medical text transcription classification with the aid of data from mtsamples.com, which provides a wide range of medical transcription samples with a focus on the different specializations. This website offers useful tools for the classification and analysis of medical texts, and it arose to meet the problems of small and restricted datasets like TCM, Hallmarks, and AIM, which have some limitations, such as a small number of classifications of medical disciplines and samples (Almazaydeh et al., 2023, p. 68). This dataset, mtsamples.csv, offers the medical transcription samples as a solution for our classification problem. This collection contains medical transcription samples from multiple specializations in the medical field. All this information was extracted from mtsamples.com. mtsamples.com was developed to provide us with convenient access to a multitude of transcribed medical record EHR data.

3.4.1 Dataset

This dataset has 4999 rows of records and six columns (‘Unnamed: 0,’ ‘description,’ ‘medical_specialty,’ ‘sample_name,’ ‘transcription,’ and ‘keywords’), as shown in Table 3 below.

Table 3. Sample data description of medical transcription dataset

	Unnamed: 0	description	medical_specialty	sample_name	transcription	keywords
0	0	A 23-year-old white female presents with comp...	Allergy / Immunology	Allergic Rhinitis	SUBJECTIVE:, This 23-year-old white female pr...	allergy / immunology, allergic rhinitis, aller...
1	1	Consult for laparoscopic gastric bypass.	Bariatrics	Laparoscopic Gastric Bypass Consult - 2	PAST MEDICAL HISTORY:, He has difficulty climb...	bariatrics, laparoscopic gastric bypass, weigh...
2	2	Consult for laparoscopic gastric bypass.	Bariatrics	Laparoscopic Gastric Bypass Consult - 1	HISTORY OF PRESENT ILLNESS: , I have seen ABC ...	bariatrics, laparoscopic gastric bypass, heart...
3	3	2-D M-Mode. Doppler.	Cardiovascular / Pulmonary	2-D Echocardiogram - 1	2-D M-MODE: , ,1. Left atrial enlargement wit...	cardiovascular / pulmonary, 2-d m-mode, dopple...
4	4	2-D Echocardiogram	Cardiovascular / Pulmonary	2-D Echocardiogram - 2	1. The left ventricular cavity size and wall ...	cardiovascular / pulmonary, 2-d, doppler, echo...

Source: Own results.

The following is an overview of the columns in the above dataset:

- Unnamed: 0: This column is added during the CSV export and is normally an index column. Fundamentally, it is not essential for analysis.
- description: A short description of the medical situation or history.
- medical_specialty: The medical specialty related to the transcription (e.g., allergy/immunology, bariatrics, cardiovascular/pulmonary, etc.).
- sample_name: The title or name of the medical sample.
- transcription: The complete transcription text of the medical sample.
- keywords: Keywords correspond with the medical sample.

Furthermore, in the above dataset structure, we are interested in three important columns: 'medical_specialty', 'transcription', and 'keywords', which are represented by Table 4 below.

Table 4. Selected 3 ('medical_specialty', 'transcription', and 'keywords') columns dataset

	transcription	keywords	medical_specialty
0	SUBJECTIVE:, This 23-year-old white female pr...	allergy / immunology, allergic rhinitis, aller...	Allergy / Immunology
1	PAST MEDICAL HISTORY:, He has difficulty climb...	bariatrics, laparoscopic gastric bypass, weigh...	Bariatrics
2	HISTORY OF PRESENT ILLNESS: , I have seen ABC ...	bariatrics, laparoscopic gastric bypass, heart...	Bariatrics
3	2-D M-MODE: , ,1. Left atrial enlargement wit...	cardiovascular / pulmonary, 2-d m-mode, dopple...	Cardiovascular / Pulmonary
4	1. The left ventricular cavity size and wall ...	cardiovascular / pulmonary, 2-d, doppler, echo...	Cardiovascular / Pulmonary

Source: Own results.

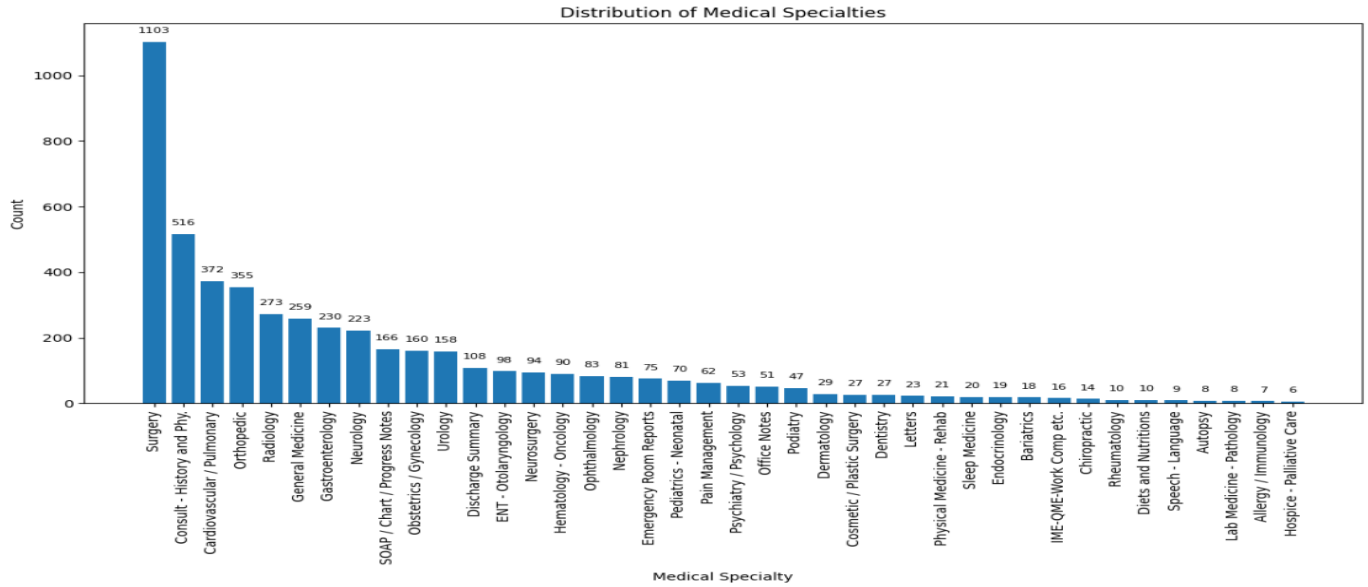
3.5 Exploratory Data Analysis (EDA)

EDA is a crucial step in the data analysis process, whose main goal is to get as much insight as possible into the structure of a given dataset. It uses both graphical and non-graphical methods to identify the correlation between variables, identify outliers and anomalies, and develop efficient models. EDA can be of two types, namely univariate EDA and multi-variate EDA, where the former focuses on a single variable while the latter deals with more than one variable at a time. Some of the main tasks include understanding the structure of the dataset, identifying remarkable patterns, and creating suitable variables for medicinal use (Komorowski et al., 2016, p. 31). We will concentrate on the dataset's three major columns: medical_specialty, transcription, and keywords.

a) Analyze categorical variable:

The first type of data visualization adopted in this univariate analysis was a bar plot that helped illustrate the extent of the medical specialization classes in the data set. Figure 21, titled "Distribution of Medical Specialties," shows the frequency of each medical specialty category. "Surgery" is the most common specialty, followed by "Consult - History and Phy" and "Cardiovascular / Pulmonary," with "Hospice - Palliative Care" ranking last. This univariate analysis provides an early grasp of the dataset's medical specialty combinations.

Figure 21. Distribution of medical specialties



Source: Own results.

b) Analyze word count by medical specialty:

"Surgery" has the most words (526,754), followed by "Consult - History and Phy" (287,961), "Orthopedic" (198,459 words), "Cardiovascular / Pulmonary" (160,867 words), and "Lab Medicine - Pathology" (1828). This analysis demonstrates a straightforward comparison of the word count of transcription columns of different medical specialties, and it is quite simple to determine the specialties with the most and least word counts in the dataset.

c) Analyze word count by keywords:

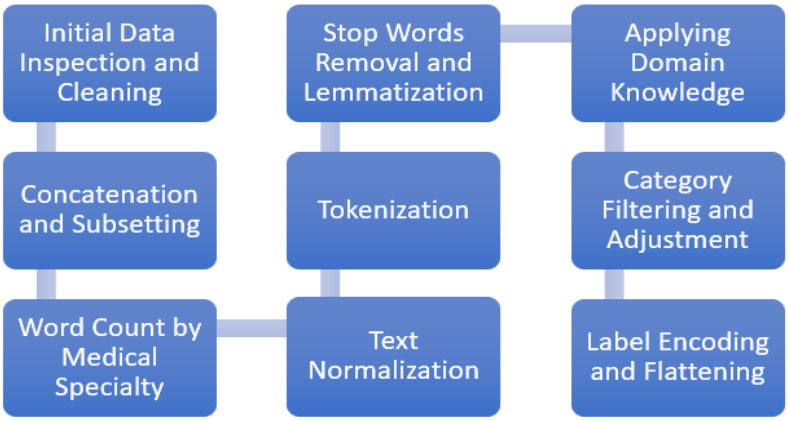
Surgery has the largest word count, with 27,858, showing a significant prevalence of various keywords. Cardiovascular/Pulmonary and Radiology have 8,278 and 6,924 word counts, respectively, indicating strong keyword usage, although at lower levels than Surgery. Interestingly, there does not seem to be a word count on Autopsy. This analysis shows that different terms are used in different specialties, which means that priorities and content in the medical sector are different.

The EDA of the dataset provides several important findings concerning medical specialties, transcription content, and keyword application. The distribution of medical specializations shows that "Surgery" is the most common, followed by "Consult - History and Phy" and "Cardiovascular / Pulmonary", while "Hospice - Palliative Care" is the least common. The word cloud illustration provided in appendix A allows for an intuitive understanding of the most common specialties. The word count investigation shows that "Surgery" has the highest transcription word count, followed by "consult - history and phy" and "orthopedic." In comparison, "Lab Medicine - Pathology" has the lowest word count. Furthermore, keyword analysis indicates that "Surgery" has the most significant keyword word count, considerably outnumbering other specialties such as "Cardiovascular/Pulmonary" and "Radiology," whereas "Autopsy" has no keyword usage. These findings emphasize the variety in content volume and keyword diversity between medical specialties, shedding light on the dataset's structure and focus regions.

3.6 Data Preprocessing

Preprocessing textual data is a critical process in natural language processing, especially in specialized domains such as medical transcription. Effective preprocessing prepares textual data for analysis, making it fit for models while enhancing its quality, which in turn enhances the models' efficiency. This procedure is divided into several stages: data inspection and cleaning, normalization, tokenization, and lemmatization. All the steps are important in reducing noise and arriving at the most appropriate data. We make sure the data is not only clean but also relevant by applying domain knowledge and transforming categories. The flowchart below in Figure 22 shows the detailed data preprocessing steps followed to transform a medical transcription dataset for analysis and modeling.

Figure 22. Data preprocessing flowchart



Source: Own representation.

3.6.1 Initial Data Inspection and Cleaning

The initial data inspection and cleaning technique is a critical stage in data preprocessing that ensures the dataset's quality and reliability. This method starts by identifying the dataset's initial structure, including its dimensions and the amount of missing information, with the goal of providing a foundation for future research. This instance includes 4999 data records and three columns. Handling missing data is a crucial part of this process, and appropriate actions are made based on the type of data. This may require removing rows with key missing values, which could have a significant impact on the results of any analysis or model generated from the data.

From the code output, it is observed that the “keywords” column has 1068 missing values while the “transcription” column has 33 missing values. The first step is to handle the missing values in the “transcription” column, which is to remove the rows with the empty “transcription” column. Also, the empty row in the “keywords” column should be empty strings. This step is important because if the transcription column is merged with the keyword’s column containing empty values, then there will be empty rows. Finally, it is essential to check the cleaning methods after correctly handling missing data. This involves checking the form of the dataset and the number of missing values after cleaning to meet the quality standards. After cleaning, the dataset has 4966 data records and three variables. By following these processes, the dataset is cleaned and made consistent for further preprocessing or analysis. This makes it possible to come up with accurate and relevant findings that will help in solving the problem.

3.6.2 Concatenation and Subsetting

In the preprocessing pipeline, the ‘keywords’ and ‘transcription’ columns were concatenated. This was done to enhance the text data by combining it with keywords that are related to the primary transcription data. Keywords often contain a lot of additional information that can help in understanding the model better. Thus, by combining these two columns, we guarantee that all the

necessary data is included in the study, which is essential when training machine learning models. After concatenation, every component in the ‘transcription’ column contains the transcription text and the linked keywords, which makes the text input more informative and detailed. After concatenation, the important columns were selected to make the analysis easier and more meaningful. The required columns ‘transcription’ and ‘medical_specialty’ were selected, and the new DataFrame is shown in Table 5. This stage is useful in reducing the dataset to the bare minimum by removing any unnecessary columns and only retaining the data that will be useful in the subsequent preprocessing and analysis stages.

Table 5. DataFrame with transcription and medical_specialty

	transcription	medical_specialty
0	allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal, erythematous, allegra, sprays, allergic,SUBJECTIVE:, This 23-year-old white female presents with complaint of allergies. She used to have allergies when she l...	Allergy / Immunology
1	bariatrics, laparoscopic gastric bypass, weight loss programs, gastric bypass, atkin's diet, weight watcher's, body weight, laparoscopic gastric, weight loss, pounds, months, weight, laparoscopic, band, loss, diets, overweight, lostPAST MEDICAL HISTOR...	Bariatrics
2	bariatrics, laparoscopic gastric bypass, heart attacks, body weight, pulmonary embolism, potential complications, sleep study, weight loss, gastric bypass, anastomosis, loss, sleep, laparoscopic, gastric, bypass, heart, pounds, weight,HISTORY OF PRESE...	Bariatrics
3	cardiovascular / pulmonary, 2-d m-mode, doppler, aortic valve, atrial enlargement, diastolic function, ejection fraction, mitral, mitral valve, pericardial effusion, pulmonary valve, regurgitation, systolic function, tricuspid, tricuspid valve, normal...	Cardiovascular / Pulmonary
4	cardiovascular / pulmonary, 2-d, doppler, echocardiogram, annular, aortic root, aortic valve, atrial, atrium, calcification, cavity, ejection fraction, mitral, obliteration, outflow, regurgitation, relaxation pattern, stenosis, systolic function, tric...	Cardiovascular / Pulmonary

Source: Own results.

The modified dataset has 4966 records in rows and two columns. By selecting only, the ‘transcription’ and 'medical_specialty’ fields, the dataset is kept concise and easy to work with, as opposed to the full dataset. This limited subset allows for proper preparation of text data processing and analysis, as well as proper classification into the right medical specialties.

3.6.3 Word Count by Medical Specialty

After concatenating the 'keywords' and 'transcription' columns and subsetting the appropriate columns ('transcription' and 'medical_specialty'), a study of the word count distribution of the transcription column across various medical specialties was performed. The findings revealed considerable differences in word count among different medical specializations. Some disciplines, such as Surgery and Consult - History and Physiology, had significantly greater word counts than others. This shows that the transcriptions for these disciplines may require more detailed or extended documentation. Specialties such as Allergy / Immunology and Lab Medicine - Pathology have lower word counts, which suggests that transcriptions in these fields could be shorter or more concise.

The relative number of transcriptions within each specialty is determined by analyzing the updated word count distribution of the transcription column by medical specialty after concatenation, which provides useful information about data distribution and potential future research or modeling directions. After combining the 'keywords' and 'transcription' columns, the total number of words in all transcriptions in the dataset was calculated. The total number of words has been found to be

2,407,470, which is useful for assessing the overall amount of content and making data handling and processing decisions.

3.6.4 Text Normalization and Tokenization

In this step, the modified transcription column's text data is normalized to enable it to be classified as medical text. This includes two major steps:

- Converting to Lowercase: This step lowercases all characters to ensure text consistency. This avoids inconsistencies in capitalization and makes it easy to compare and understand the information (Kowsari et al., 2019, p. 4).
- Removing Punctuation and Numbers: In text analysis tasks, punctuation and numbers may even turn out to be a source of noise and may be dismissed sometimes. Thus, by excluding them, the language becomes more concise as the specifically important words and phrases dominate. This stage consists of the use of regular expressions, which are used to filter out punctuation and numbers from the text (Pahwa et al., 2018, p.16).

Normalizing the text in these ways improves the data's clarity, consistency, and suitability for tasks such as text categorization. Tokenization is the next step following text normalization. Tokenization is the process of breaking down a text into smaller parts, typically words or tokens, in order to facilitate further analysis. This approach helps to turn the text into a more understandable structure for computational analysis (Rai & Borah, 2021, p. 194). As a result, tokenization is a key preprocessing step in natural language processing that enables efficient text analysis and modeling.

3.6.5 Stop Words Removal and Lemmatization

Following tokenization and stemming, the text is often further refined by eliminating stopwords and implementing lemmatization.

- Stop word Removal: A stop word is a word or phrase that does not carry any meaning in a language and can be eliminated without having a negative impact on the result of many related operations. The deletion of stop words from the text enables the model to give a precise identification of the valuable terminologies in the text (Sarica & Luo, 2021, p. 1). This stage includes filtering out stop words from the tokenized text.
- Lemmatization: Lemmatization is the process of reducing words to their basic or dictionary form, i.e., a lemma. Lemmatization precisely reduces words to their root form using context and morphological analysis, compared to stemming, which just eliminates suffixes. This step ensures that the many inflected versions of a word are treated as one. This may improve the coherence and validity of the analysis (Di Nunzio & Vezzani, 2018, p. 182).

The text is cleaned from stop words and lemmatized to become more proper for tasks like text categorization. They play the roles of noise minimization, feature augmentation, and quality, in addition to improving performance in natural language processing. Furthermore, it is worth providing the

following Table 6, which shows the impact of all these pre-processing steps on the given example of an input text.

Table 6. The effect of pre-processing steps on a sample input text

Pre-Processing Steps	Effect of Pre-Processing Steps on a Sample Input Text
Input text	allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal, erythematous, allegra, sprays, allergic,SUBJECTIVE:, This 23-year-old white female presents with complaint of allergies.
Text Lowering	allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal, erythematous, allegra, sprays, allergic,subjective:, this 23-year-old white female presents with complaint of allergies.
Removing Punctuation and Numbers	allergy immunology allergic rhinitis allergies asthma nasal sprays rhinitis nasal erythematous allegra sprays allergicsubjective this yearold white female presents with complaint of allergies
Tokenization	allergy, immunology, allergic, rhinitis, allergies, asthma, nasal, sprays, rhinitis, nasal, erythematous, allegra, sprays, allergicsubjective, this, yearold, white, female, presents, with, complaint, of, allergies,
Stop word removal	allergy, immunology, allergic, rhinitis, allergies, asthma, nasal, sprays, rhinitis, nasal, erythematous, allegra, sprays, allergicsubjective, yearold, white, female, presents, complaint, allergies, used, allergies,
Lemmatization	allergy, immunology, allergic, rhinitis, allergy, asthma, nasal, spray, rhinitis, nasal, erythematous, allegra, spray, allergicsubjective, yearold, white, female, present, complaint, allergy, used, allergy,

Source: Own results.

3.6.6 Applying Domain Knowledge

Following the removal of stopwords and lemmatization, domain knowledge is used to refine and adapt the dataset's categories, which is especially useful in medical text classification applications. Categories such as 'Surgery', 'SOAP/ Chart / Progress Notes', 'Office Notes', 'Consult - History and Phy.', 'Emergency Room Reports', 'Discharge Summary', 'Pain Management', 'General Medicine', and 'Letters' are excluded because they are either too general or not specific to medical specializations. Additionally, modifications are made to ensure that the produced dataset meets the desired numbers for medical specialties; for instance, ‘Neurosurgery’ cases are reassigned to ‘Neurology’ whereas ‘Nephrology’ instances go to ‘Urology’. This technique means that the dataset is fine-tuned

for the specifics of medical-related text and, as a result, makes the subsequent classification-related tasks more relevant and precise.

Following these modifications, the dataset is provided in more appropriate groups regarding the newly enhanced medical specialties, which generally makes great sense in medical text categorization. Grouping provides a better understanding of the distribution of samples across different medical disciplines, hence aiding in the formulation of better classification models. It is for this reason that we can train classifiers to differentiate between various medical specializations when the data is arranged in this systematic manner; the training will be useful in discovering medical disorders and patient record routing. By using domain knowledge and after redefining the categories, the dataset, which originally contained 2,620 rows of records and two columns, has reduced the medical specialty count to 29 from 40. Notably, Cardiovascular / Pulmonary has the largest sample count (371), followed by Orthopedic (355) and Neurology (317). Hospice - Palliative Care had the fewest samples, with only six.

3.6.7 Category Filtering and Adjustment

By incorporating medical domain knowledge and changing the categories, the next steps included in the refinement of the dataset involved filtering and thus adjusting the categories. Given the wide range in sample counts between medical specialties, categories with fewer than 50 samples were excluded in order to concentrate on those with a larger representation. This stage makes sure the dataset includes categories with sufficient data for robust analysis and model training. The resultant data set now has 2,324 rows and two columns ("preprocessed_transcription" and "medical_specialty"), with each transcription record corresponding to a specific medical specialty. Table 7 below provides the most common medical specialties and their counts.

Table 7. Finalized medical specialty and respective count

Medical Specialty	Count
Cardiovascular / Pulmonary	371
Orthopedic	355
Neurology	317
Radiology	273
Urology	237
Gastroenterology	224
Obstetrics / Gynecology	155
ENT - Otolaryngology	96
Hematology - Oncology	90
Ophthalmology	83
Pediatrics - Neonatal	70
Psychiatry / Psychology	53

Source: Own results.

This filtering and modification step is critical in ensuring that the given set of data is balanced and well representative of the actual world medical classes, thus enhancing the reliability as well as effectiveness of the following studies and classification models. In that way, focusing on the categories that provide enough data can improve the understanding of some medical specializations and develop better predictive models for the tasks connected with medical text categorization. In addition, the transcription column comprised 2,407,470 words prior to data preprocessing. The word counts drastically decreased to 629262 words after using various preprocessing techniques such as removing punctuation, stopwords, and lemmatization. This reduces the total word count by 73.86%.

3.6.8 Label Encoding and Flattening

Upon incorporating medical domain knowledge and altering the categories, the dataset was refined further by category filtering and adjustment. The next step in the procedure is to encode the medical specialties' categorical classifications into numerical values that machine learning algorithms can understand. This transformation produces a new column with these numerical labels, making the data more machine-learning model-friendly. This step is critical since many machine learning methods require numerical input to work properly, and encoding the category data guarantees that the medical specialties are represented in a format that the models can understand. Following label encoding, the text data is converted into a vectorization-ready format. The normalized and cleaned transcription data is organized in a nested list structure. This format is not suitable for direct input into vectorization procedures, so the text data is flattened into a single list. This flattened list of transcriptions can be easily fed into the text vectorization processes, which convert textual data into numerical values. These features are now indispensable for training machine learning models, enabling algorithms to learn from text data and classify and predict based on medical text. The preprocessed dataset that is ready for model building is presented in Table 8 with 2324 data records and three fields.

Table 8. Final preprocessed medical transcription dataset

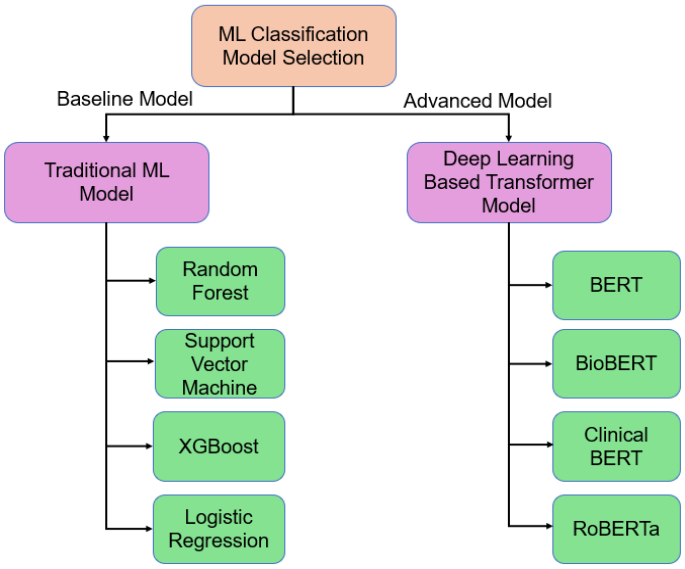
	preprocessed_transcription	medical_specialty	encoded_target
3	cardiovascular, pulmonary, mmode, doppler, aortic, valve, atrial, enlargement, diastolic, function, ejection, fraction, mitral, mitral, valve, pericardial, effusion, pulmonary, valve, regurgitation, systolic, function, tricuspid, tricuspid, valve, nor...	Cardiovascular / Pulmonary	0
4	cardiovascular, pulmonary, doppler, echocardiogram, annular, aortic, root, aortic, valve, atrial, atrium, calcification, cavity, ejection, fraction, mitral, obliteration, outflow, regurgitation, relaxation, pattern, stenosis, systolic, function, tricu...	Cardiovascular / Pulmonary	0
7	cardiovascular, pulmonary, echocardiogram, cardiac, function, doppler, echocardiogram, multiple, view, aortic, valve, coronary, artery, descending, aorta, great, vessel, heart, hypertrophy, interatrial, septum, intracardiac, pericardial, effusion, tri...	Cardiovascular / Pulmonary	0
9	cardiovascular, pulmonary, ejection, fraction, lv, systolic, function, cardiac, chamber, regurgitation, tricuspid, normal, lv, systolic, function, normal, lv, systolic, ejection, fraction, estimated, normal, lv, lv, systolic, systolic, function, funct...	Cardiovascular / Pulmonary	0
11	cardiovascular, pulmonary, study, doppler, tricuspid, regurgitation, heart, pressure, stenosis, ventricular, heart, ventricle, tricuspid, regurgitationd, study, mild, aortic, stenosis, widely, calcified, minimally, restricted, mild, left, ventricular,...	Cardiovascular / Pulmonary	0

Source: Own results.

3.7 Model Selection

In ML and AI, choosing suitable models is the key to the success of any predicting or analytic undertaking. Model selection is the process of identifying the right methodology and structure to extract the underlying patterns and dependencies of the data for the purpose of making correct forecasts and providing valuable information. In this regard, the model selection procedure is strictly regulated to handle the complexity and hierarchy characteristic of classification tasks; our methods encompass classical machine learning techniques together with modern deep learning algorithms, as represented in Figure 23 below.

Figure 23. Proposed methodology for ML medical text classification model selection



Source: Own representation.

The first step in our model selection process is to look at the basic machine learning methods that are also referred to as the baseline models. These models are the fundamental ones, which give a benchmark by which other, more complex methods can be compared. In this discipline, we have chosen algorithms that are characterized by their flexibility, robustness, and interpretability.

3.7.1 Baseline Models (Traditional ML)

Our baseline models are built around the following (Ozcan et al., 2022, pp. 3289-3290):

- Random Forest: Random Forest is based on decision trees and is capable of handling high dimensions; therefore, it is suitable for cases with complex relations and noise.
- Support Vector Machine: SVM is particularly effective in classifying data in high-dimensional spaces, which is always useful in applications with complex decision planes based on the hyperplane separation concept.
- XGBoost: XGBoost is a gradient-boosting algorithm that is well-known for its scalability and performance. XGBoost is best suited for structured or tabular data.

- Logistic Regression: Despite its simplicity, logistic regression still remains a strong technique, particularly when it comes to binary classification problems that require both interpretable and fast algorithms.

3.7.2 Advanced Models (Deep Learning Based Transformer Models)

The choice of models does not stop at traditional approaches, and our model selection is based on the revolutionary potential of deep learning, as transformer models show. These sophisticated models, which are often defined by the capability of recognizing intricate patterns from huge amounts of data, are the current state of the art in AI-based categorization.

- BERT: BERT (version ‘bert-base-uncased’) is making a paradigm shift in natural language processing by utilizing both the forward and backward context and training on extensive text data to achieve the best results in most of the NLP tasks (Devlin et al., 2019, p. 4172).
- BioBERT: BioBERT, which is designed for the field of biological text processing, enriches the BERT’s functionality by adding domain knowledge and semantics to make breakthroughs in biomedicine and healthcare (Lee et al., 2020, p. 1235). The version of BioBERT used in this project is 'dmis-lab/biobert-v1.1'.
- Clinical BERT: Clinical BERT (version 'emilyalsentzer/Bio_ClinicalBERT') is specifically used to work with medical concepts and terminologies, revealing information essential for clinical decision-making and research. It is concerned with the inherent complexity of clinical content typical of EHRs and medical documents (Alsentzer et al., 2019, p. 73).
- RoBERTa (Robustly optimized BERT approach): RoBERTa (version 'roberta-base') builds on the groundwork laid by BERT, which includes modifications in the training process and hyperparameters and shows enhanced performance and stability in different NLP tasks (Liu et al., 2019, p. 2).

3.8 Model Development and Evaluation

In the model development of machine learning techniques, it is the process of developing and fine-tuning models with a view to identifying patterns that exist in the data and the correlation between them. Therefore, the methods that are most suitable for the given task are chosen, data is prepared, and hyperparameters are adjusted to get the best results. For their evaluation, models are trained with the training data and evaluated with validation datasets. The experimentation and assessment procedure cannot be completed once; it must be repeated on a regular basis to improve the model's accuracy, generalization, and resilience. This is to develop a model that is credible and efficient for making predictions on new and unforeseen data.

3.8.1 Baseline Model Development

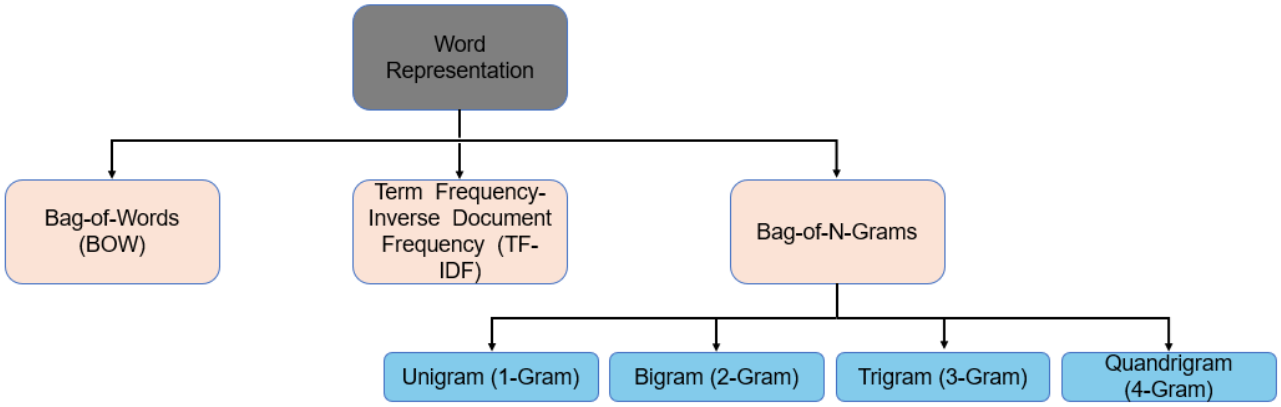
As discussed in section 3.7, we begin by building a baseline model, as shown in Figure 24.

Figure 24. Proposed method for the baseline model development

Source: Own representation.

In this method, we must start with the word representation phase, which comprises transforming text input into some vector representation so that the algorithms can automatically understand analogies and generalize that term. Word representation methods range from classical to representation learning (Naseem & Musial, 2019, p. 953). For the medical text categorization task, we utilized three popular models, as shown in Figure 25 below. The models are described below.

Figure 25. Three common word representation methods adopted for medical text classification task



Source: Own representation.

a) Word representation

- Bag-of-Words (BOW): The BOW method quantifies text by the number of words used in different texts. It maps text to numerical features for the algorithms to process and analyze the text data. BOW, on the other hand, does not consider the position of the words in the text or the context of the words used but only the occurrence of the word. It is convenient and efficient for dealing with text, but it can be disadvantageous in that it loses some of the meaning nuances (Juluru et al., 2021, p. 1424).
- Term Frequency-Inverse Document Frequency (TF-IDF): Term Frequency-Inverse Text Frequency is one of the feature extraction weighting factors that measures the importance of terms in a text within the corpus. It uses term frequency (TF), which is the number of times a term

occurs in a document, with inverse document frequency (IDF), which measures the rarity of the terms in the given collection. The measure proposed in this work helps to find words that are frequent in a document but rare in the collection, which is useful for tasks such as text classification. TF-IDF is essential in information retrieval as well as text mining for identifying the importance of phrases (Das & Chakraborty, 2018, p. 3).

- Bag-of-n-grams: The bag-of-n-grams model is one of the most commonly used techniques to represent text data with n-gram characteristics. This approach refers to the text by counting or calculating the frequency of n-grams, or even using a bag of words model where the order and context of the n-grams are not considered. This representation can be used as input for most machine learning algorithms, like SVMs, logistic regression, and neural networks, for tasks like text classification (B. Li et al., 2016, p. 1592). For an example of feature extraction using n-grams, let the text be “The student is alone happily.” The total number of n-gram features can be found as k-n+1, where k is the number of words. The final outcome is the bag-of-n-grams model for a classifier, which will train the linguistic algorithm (Marafino et al., 2014, p. 872). Table 9 below demonstrates alternative n-gram feature representations formats.

Table 9. Demonstration of different n-gram feature representations

N-gram	N-gram Generated Sentence	Number of N-gram Features
Unigram (1-Gram)	“The”, “student”, “is”, “alone”, “happily”	5
Bigram (2-Gram)	“The student”, “student is”, “is alone”, “alone happily”	4
Trigram (3-Gram)	“The student is”, “student is alone”, “is alone happily”	3
Quadrigram (4-Gram)	“The student is alone”, “student is alone happily”	2

Source: Adapted from, Win & Hoon, 2022, p. 1.

b) Dimension reduction using PCA

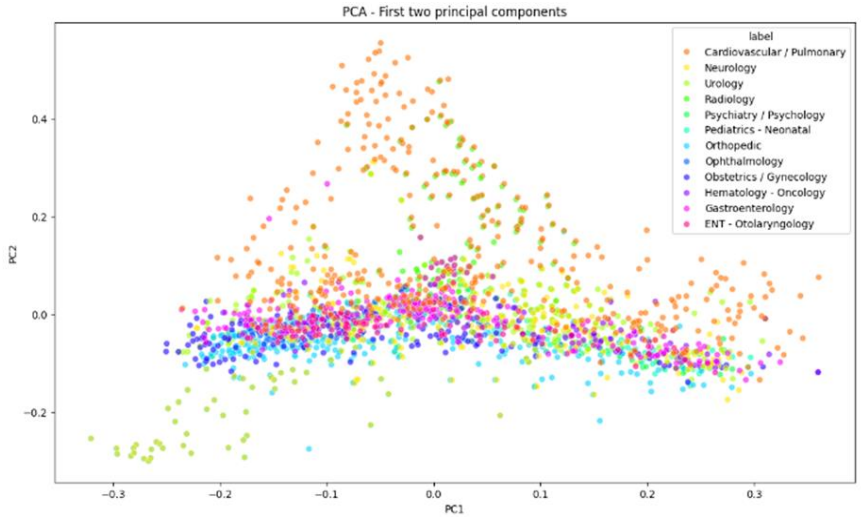
Principal component analysis (PCA) is a technique for transforming high-dimensional data into a smaller dimension while keeping crucial information. It decreases variation by transforming initial features into new linear combinations of those features, making data analysis and plotting easier. The success rate of PCA is determined by the percentage of total variance that is retained, which makes PCA a universal tool for modeling patterns in large sets of data (Jolliffe & Cadima, 2016, pp. 2-3). In this study, we explain PCA implementation using the example of a TF-IDF word representation. In this case, the PCA implementation will be described using the example of a TF-IDF word representation. The process of PCA is then performed on the input data to reduce its dimensionality to only retain 95 percent of the variance. This is done by setting n_components = 0.95 when

initializing PCA. The modified data has fewer dimensions but contains most of the elements of the original data.

i. Visualizing PCA results

Following PCA, the transformed data is organized into a DataFrame, with separate rows representing different data points and columns delineating the principal components. Each row in this data frame contains the reduced-dimensional representation of a single data point. Furthermore, this data frame includes relevant label information that indicates the classification of every data point. A scatter plot has been produced using the first two principal components as axes, with one dictating the x-coordinate and the other the y-coordinate. This scatter plot, known as Figure 26, is a visual tool for identifying probable trends, patterns, or clusters within the reduced-dimensional space. The scatter plot improves its capacity to explain by coloring data points according to their labels, making it easier to identify and analyze any observable structures.

Figure 26. PCA scatter plot of first two principal components (PC1 and PC2)



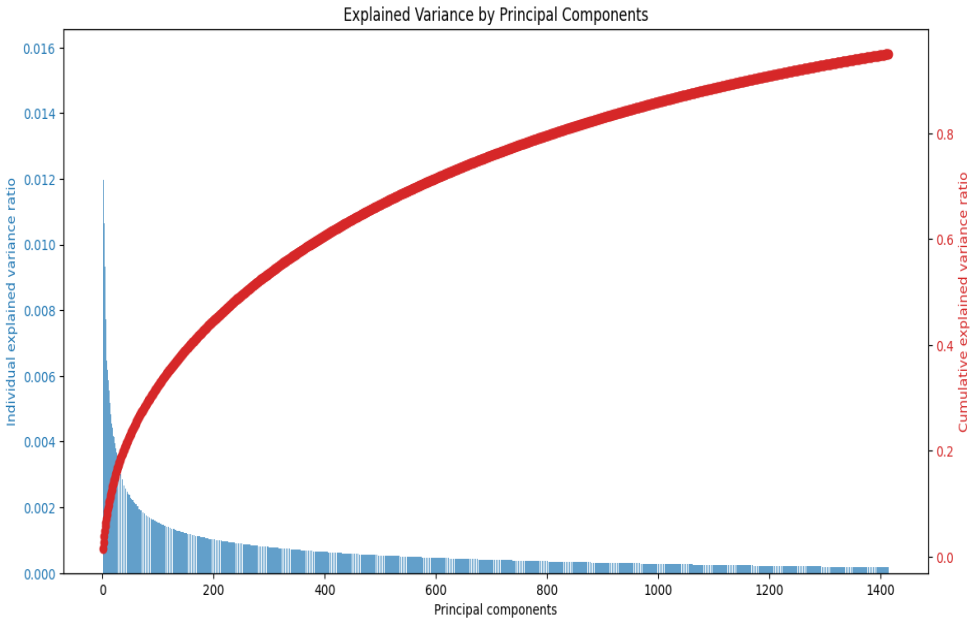
Source: Own results.

ii. Plotting Explained Variance

In PCA, the technique used determines the explained variance for each of the principal components, and this gives information on the importance of the components in preserving information from the original dataset. This data is presented in the form of graphs. The graph for “Individual Explained Variance Ratio” demonstrates that the explained variance ratio of the component increases with the sequential components, and the initial components have more variance than the subsequent components. Meanwhile, the "Cumulative Explained Variance Ratio" bar graph depicts the increase in variance explained when more components are added. Beginning from point 0, the bars go up steeply to depict a fast increase but fall rapidly, illustrating that more components contribute less to the total variance, as illustrated in Figure 27. Together, these visualizations aid in understanding the

trade-offs associated with dimensionality reduction and provide valuable insights into the entire data structure.

Figure 27. Variance by principal components explained by individual and cumulative variance ratio



Source: Own results.

c) Classification

Following text preprocessing, word representation, and dimension reduction, we can continue to the classification task. The data is imbalanced, so we must deal with it in this work using the SMOTE method. To solve imbalanced datasets, SMOTE (Synthetic Minority Over-sampling Technique) generates synthetic examples of the minority class. It discovers minority class instances, locates their nearest neighbors, and then interpolates between them to generate new synthetic cases. This strategy aids in the balance of class distribution, thereby reducing overfitting and bias in machine learning models. SMOTE can prove helpful for enhancing model performance, especially for classifiers that are sensitive to class distribution. However, effectiveness should be assessed in terms of the unique dataset and problem domain (Elreedy & Atiya, 2019, p. 34). In the following step of baseline model development, we have chosen a data split strategy of 80% train and 20% test. After splitting, a pipeline was utilized to implement the classifiers. Pipelining is employed to improve the flow of an algorithm. The multiple classifiers described in model selection section 3.7.1 for medical text categorization are used.

d) Model evaluation and result

In this investigation, four classification performance evaluation metrics were employed, namely accuracy, precision, recall, and F1 score. These assessments have four possible outcomes: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The performance

evaluation metrics with equations for four classification performance measures are listed below (Sokolova & Lapalme, 2009, p. 430).

i. Precision: Precision is defined as the ratio of true positives to all predicted positives. It determines the accuracy of positive predictions.

Precision=$\frac{TP}{(TP+FP)}$

Macro Average Precision= $\frac{1}{n}\sum_{i=1}^{n} Precision_i$

This represents the average precision across all classes. It treats all classes equally, regardless of their size.

ii. Recall: Recall is the ratio of true positives to total positives (true positives plus false negatives). It evaluates a model's capability to capture all relevant occurrences.

Recall=$\frac{TP}{(TP+FN)}$

Macro Average Recall= $\frac{1}{n}\sum_{i=1}^{n} Recall_i$

This is the average recall for all classes, with the algorithm's performance improving as it gets closer to 1. It also delivers the same amount of attention to all classes, regardless of size.

iii. F1 Score: The F1 score is determined as the harmonic average of precision and recall. It balances the two measures and is effective when the class distribution is imbalanced.

F1 Score=2* $\frac{(Precision*Recall)}{(Precision+Recall)}$

Macro Average F1 Score=$\frac{1}{n}\sum_{i=1}^{n} F1\ Score_i$

This is the mean F1 score for all classes. It provides a single performance measure that equally weighs precision and recall for each class.

iv. Accuracy: Accuracy is the model's overall correctness. It denotes the proportion of correctly classified occurrences (including true positives and true negatives) to the total number of instances.

Accuracy= $\frac{(TP+TN)}{(TP+FP+FN+TN)}$

Precision, recall, and F1 score are all averaged; however, accuracy is not. Instead, it is calculated from the entire dataset.

Accuracy= $\frac{\sum_{i=1}^{n}(TP_i+TN_i)}{\sum_{i=1}^{n}(TP_i+FP_i+FN_i+TN_i)}$

Furthermore, the performance across multiple ML techniques for classification and word representation models is shown in Table 10.

Table 10. Performance of the different classifiers and word representation models

Classifier Model	Word Representation	Macro Average Precision	Macro Average Recall	Macro Average F1 Score	Accuracy
Random Forest	BOW	0.75	0.75	0.75	0.74
	TF-IDF	**0.77**	**0.78**	**0.78**	**0.77**
	BON-gram (1,1)	0.74	0.75	0.74	0.74
	BON-gram (1,2)	0.74	0.75	0.74	0.74
	BON-gram (2,2)	0.69	0.70	0.69	0.69
	BON-gram (2,3)	0.69	0.70	0.69	0.69
	BON-gram (3,3)	0.57	0.58	0.56	0.56
Support Vector Machine	BOW	0.80	0.79	0.79	0.78
	TF-IDF	**0.80**	**0.81**	**0.80**	**0.80**
	BON-gram (1,1)	0.77	0.76	0.76	0.75
	BON-gram (1,2)	0.76	0.76	0.75	0.75
	BON-gram (2,2)	0.61	0.54	0.54	0.53
	BON-gram (2,3)	0.57	0.49	0.49	0.47
	BON-gram (3,3)	0.42	0.31	0.30	0.31
XGBoost	BOW	0.75	0.75	0.75	0.74
	TF-IDF	**0.82**	**0.82**	**0.82**	**0.82**
	BON-gram (1,1)	0.75	0.75	0.75	0.74
	BON-gram (1,2)	0.75	0.75	0.75	0.74
	BON-gram (2,2)	0.71	0.71	0.71	0.70
	BON-gram (2,3)	0.70	0.71	0.70	0.70
	BON-gram (3,3)	0.58	0.59	0.58	0.58
Logistic Regression	BOW	0.80	0.80	0.80	0.79
	TF-IDF	**0.82**	**0.82**	**0.82**	**0.81**
	BON-gram (1,1)	0.80	0.80	0.80	0.80
	BON-gram (1,2)	0.79	0.80	0.79	0.79
	BON-gram (2,2)	0.69	0.69	0.68	0.68
	BON-gram (2,3)	0.66	0.66	0.65	0.65
	BON-gram (3,3)	0.51	0.47	0.46	0.46

Source: Own results.

Based on the findings in Table 10, we can gauge the F1 score, which balances precision and recall. Considering the F1 scores for various classifiers and word representation methods:

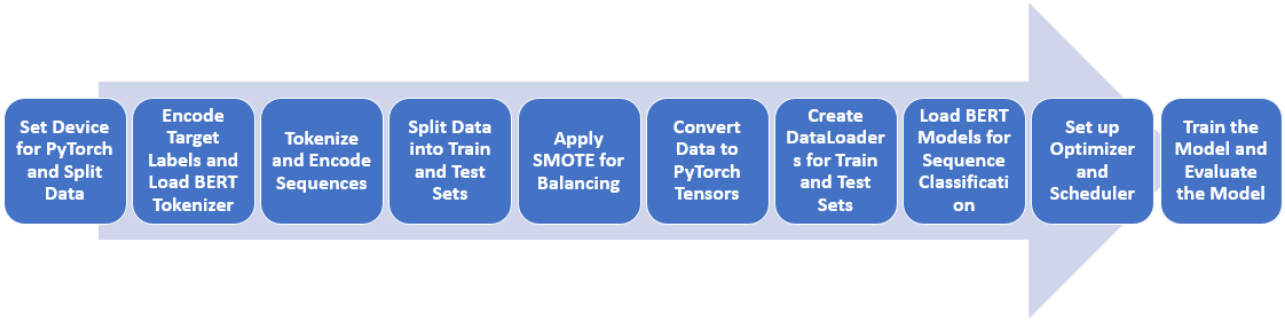
- TF-IDF representation has the highest F1 score of 0.78 in random forest analysis.
- The TF-IDF representation achieves the highest F1 score of 0.80 in the support vector machine.
- The TF-IDF representation has the best F1 score for XGBoost at 0.82.
- The TF-IDF representation achieves the highest F1 score of 0.82 in logistic regression.

According to these outcomes, the TF-IDF representation regularly outperforms all other classifiers. As a result, TF-IDF is likely the best choice for the word representation method in this circumstance; hence, it has been selected as the base model for comparison with the advanced model.

3.8.2 Advanced Model Development

As we discussed in section 3.7, after developing the baseline model, we must now develop the advanced model, as illustrated by Figure 28 below.

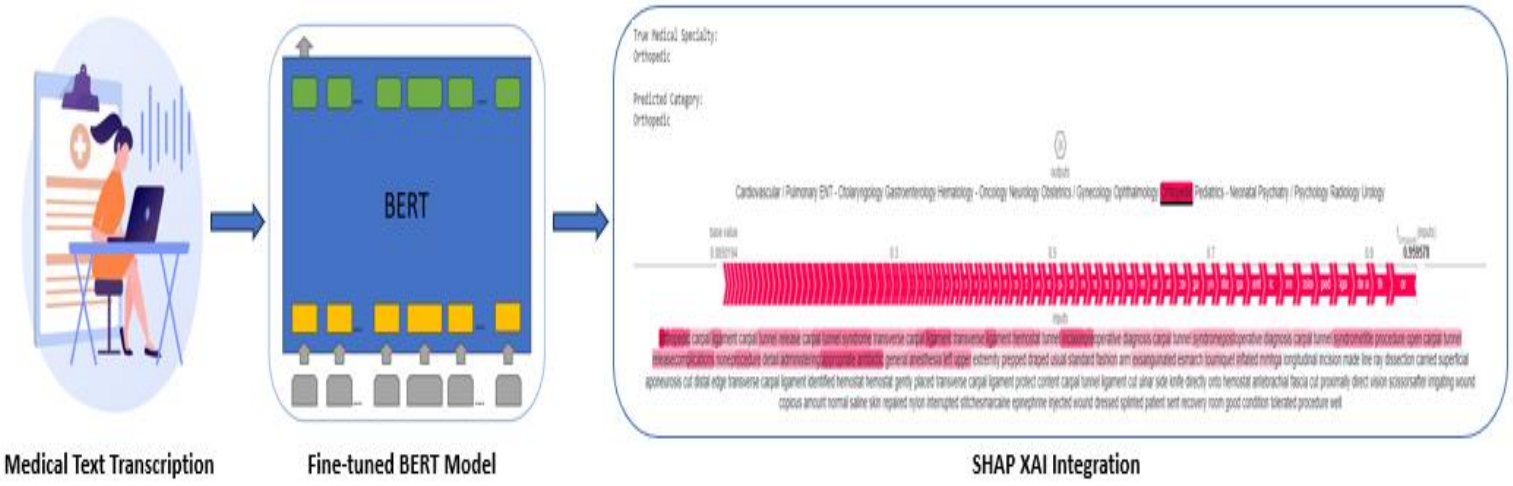
Figure 28. Proposed method for the advance model development



Source: Own representation.

Moreover, Figure 29 shows a proposed system in which physician notes and medical transcripts provide input for a model. The model generates a medical specialty classification with color coding that shows the relevance of terms in determining the model's decision: red indicates the most important phrases that favorably impact the prediction, blue indicates the words that negatively influence the prediction, and white indicates a neutral influence. This approach aims to provide an understanding of the model's decision-making process, as well as a more accurate and convenient way of selecting a relevant medical specialty. The proposed method, which includes an explainability component, may help to increase confidence in the application of machine learning for categorization of medical text transcription.

Figure 29. A proposed system of using deep learning transformer-based BERT as advanced model with XAI technique for decision making



Source: Own representation.

The below steps describe the proposed method for the advance model development:

i. Set device for PyTorch and split data: This stage first checks if there is a CUDA-enabled GPU available; if yes, then it sets the device to CUDA; otherwise, it sets it to CPU. We are executing the code on a Google Collab notebook, for which they have an NVIDIA A100 GPU available. This option allows PyTorch to use GPU acceleration, which can dramatically speed up the training process. The data is then separated into features (X) and targets (Y).

ii. Encode target labels and load the BERT tokenizer: Utilizing label encoding, convert the category target labels (medical_specialty) to a numerical format. Load the BERT tokenizer, which is required for transforming textual data into a format that the BERT model can understand.

iii. Tokenize and encode sequences: To tokenize and encode textual sequences into numerical tokens, utilize the loaded BERT tokenizer. This stage prepares the data for feeding into the neural network model.

iv. Split data into train and test sets: To appropriately train and evaluate the model, we split the data into training and testing subsets. We allocate 80% of the data to the training set and 20% to the testing set. The training set trains the model, and the testing set evaluates its performance on previously unknown data.

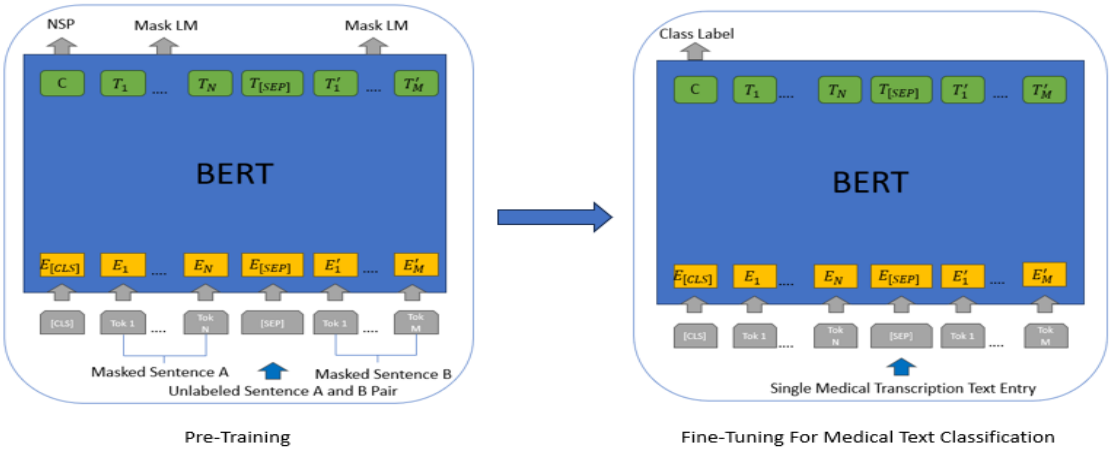
v. Apply SMOTE for balancing: To deal with the class imbalance in the advanced modeling, use SMOTE on the training dataset. SMOTE creates synthetic samples for minority classes with the aim of improving the class distribution of the dataset.

vi. Convert data to PyTorch tensors: Convert the tokenized and encoded data to PyTorch tensors. Tensors can be described as high-dimensional arrays that have been specifically designed for use in deep learning modeling tasks.

vii. Create DataLoaders for train and test datasets: To load and process data in a quick way in the training and testing phases, create DataLoaders. DataLoaders are involved in handling activities like batching, shuffling, and concurrent processing so that the model training is seamless.

viii. Load BERT models for sequence classification: Load a pre-trained BERT model for sequence classification tasks, as shown in Figure 30, followed by fine-tuning for medical text categorization. BERT is an effective language model that specializes in understanding contextual relationships within textual input. (Left) The original pre-trained BERT is trained to do 'next sentence prediction (NSP)' and 'masked-language modeling (MLM)'. To facilitate learning, specific categorization [CLS] and separator [SEP] tokens are added to the input. (Right) BERT has been fine-tuned for this medical text classification task by analyzing labeled data from medical transcript physician entries. The result is a class label associated with the assigned medical specialization.

Figure 30. The pretrained BERT model and the fine-tuning for medical text classification



Source: Adopted from: Talebi et al., 2024, p. 4.

ix. Set up optimizer and scheduler: The BERT model will be fine-tuned with an AdamW optimizer that has a learning rate of 2 x 10^{-5} and an epsilon value of 1 x 10^{-8}. Further, we will use a linear learning rate scheduler with no warm-up steps, which means that the learning rate will be modified consistently even in the initial epochs. The current model will be trained for four epochs to reintroduce the set of first-level key frame identifiers to improve parameter estimates. These components allow for adjustments to the model and learning rates during training, and thus efficient and useful training is achieved.

x. Train the model and evaluate the model: For the loss function, the BERT model needs to be trained via available training data and model parameters' tuning. Subsequently and on a continuous basis, tests use the testing data to determine the accuracy and efficiency of the model while handling new data that has not been experienced before. This step shows how optimally the model was trained and if it is capable of predicting new sets of data that it has not encountered. Hence, we evaluate the classification performance of the developed model using four parameters as follows: precision, recall, F1 score, and accuracy, which are also used for evaluating the baseline model. The BioBERT, ClinicalBERT, and RoBERTa medical text categorization algorithm models' performance is given in Table 11.

Table 11. Summarized results of the performance of the different BERT classifier models

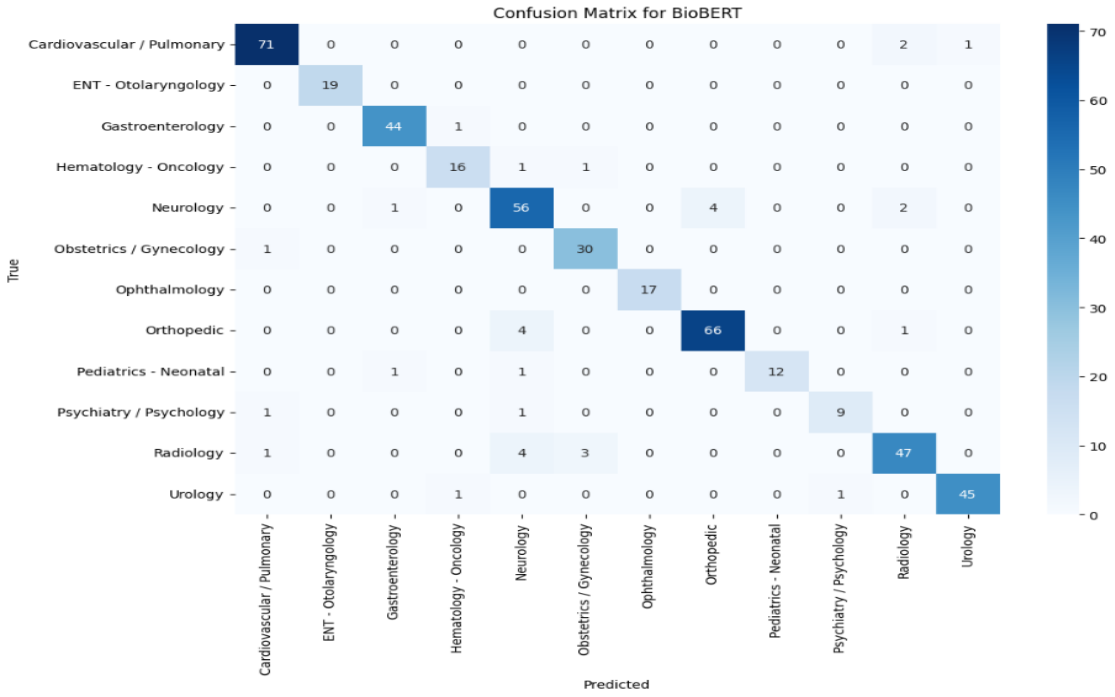
BERT Classifier Model	Macro Average Precision	Macro Average Recall	Macro Average F1 Score	Accuracy
BERT	0.92	0.88	0.90	0.90
BioBERT	0.94	0.91	0.92	0.93
ClinicalBERT	0.92	0.90	0.91	0.92
RoBERTa	0.92	0.91	0.92	0.92

Source: Own results.

According to the above-mentioned results, BioBERT is the best-performing model among the specified classifiers. It has the highest macro-average precision of 0.94 and a tie with RoBERTa for the highest macro-average recall of 0.91. BioBERT also has the highest macro-average F1 score of 0.92, which equals RoBERTa. It performs outstandingly better than the other models to a level of 0.93 on accuracy through the test datasets. These better performance measures suggest that BioBERT is especially well-suited to applications requiring high precision, recall, and overall classification accuracy. As a result, BioBERT is the best option for applications that require robust and reliable medical text classification. Furthermore, the following model metrics, such as the confusion matrix, K-fold cross-validation, and misclassification error analysis, provide a more comprehensive assessment of our BioBERT model's reliability and performance.

a) Confusion Matrix and Evaluation: A confusion matrix is a table in machine learning used for the assessment of the performance of a classification model against the actual outcomes. It reveals how often each class is classified correctly and how often each is misclassified, which aids in determining which classes are being confused (Ting, 2010, p. 347). To present the confusion matrix of the optimal choice BioBERT model, Figure 31 depicts the matrix in the form of a heatmap; thus, the intensity of the blue color represents the frequency of the occurrence. The number of predictions is shown in annotations within each cell, and the axes are named 'Predicted' on the x-axis and 'True' on the y-axis to ensure that the meaning of the rows and columns is clear. The plot is given the name "Confusion Matrix for BioBERT" to provide context information on the model being assessed. This graphic is useful in assessing the accuracy of the model by showing which predictions were true or false.

Figure 31. Confusion matrix for BioBERT medical text classification model



Source: Own results.

b) K-Fold Cross-Validation: K-fold cross-validation (CV) evaluates the model by dividing the dataset into K sets, then using one set for testing and the other K-1 sets for training. This method decreases the variance and bias of the performance estimate, providing a more robust evaluation than a single train-test split (White & Power, 2023, pp. 2-3). The K-Fold Cross-Validation findings with k = 5 indicate the BioBERT model's strong performance across many folds, demonstrating its reliability and uniformity in dealing with varied subsets of medical text data. By averaging metrics across folds, we reduce the variability that can occur from a single train-test split, resulting in a more trustworthy assessment of the model's generalization ability. The average precision, recall, and F1-score metrics are consistently near 92%, and a tabular representation of the macro and weighted average metrics across all folds based on the medical text classification reports is shown in Table 12 below, indicating strong performance across various medical specialties and highlighting the model's ability to generalize well to unseen data. This shows that BioBERT effectively learns meaningful patterns from the dataset, which is critical for its use in real-world applications requiring accuracy and robustness.

Table 12. Macro and weighted average classification metrics across 5 folds (K=5) for medical text classification report

	Precision	**Recall**	**F1-score**
Macro Average	0.920	0.920	0.920
Weighted Average	0.924	0.923	0.922

Source: Own results.

In addition, the application of SMOTE during the initial steps helps in reducing the class imbalance, which can be useful in enhancing the performance of the model across the board for all the medical specialties. This makes the presented model more trustworthy, as the training in each of the folds is stable, which means that there is constant loss reduction and that optimalization strategies such as gradient clipping and learning rate scheduling are well applied. These findings not only confirm the BioBERT model's capability for deployment but also provide actionable information for future developments, such as focused improvements in specific medical specialties or changes to training methodologies that will increase performance metrics.

c) Misclassification Error Analysis of BioBERT Model: Misclassification error analysis is an important stage in determining the performance of our BioBERT classification model. It entails looking at individual instances when the model's predictions do not match the actual labels in order to uncover trends or common traits among misclassified instances. The findings of this study make it possible to determine the strengths and shortcomings of the model by identifying whether certain kinds of input data often lead to incorrect predictions. With these understandings, it will be possible to guide further improvement of the training process of the model and data preprocessing to make our BioBERT more reliable and effective.

Table 13. Misclassification error analysis of BioBERT model

Medical Specialties	Number of Entries	Number of BioBERT Misclassifications Errors	Accuracy (%)
Cardiovascular / Pulmonary	74	3	95.94
ENT - Otolaryngology	19	0	100.0
Gastroenterology	45	1	97.77
Hematology - Oncology	18	2	88.88
Neurology	63	7	88.88
Obstetrics / Gynecology	31	1	96.77
Ophthalmology	17	0	100.0
Orthopedic	71	5	92.95
Pediatrics - Neonatal	14	2	85.71
Psychiatry / Psychology	11	2	81.81
Radiology	55	8	85.45
Urology	47	2	95.74

Source: Own results.

According to the misclassification analysis in Table 13, classes such as ENT - Otolaryngology and Ophthalmology have perfect accuracy, while Gastroenterology also demonstrates almost high accuracy, indicating robust model performance. In contrast, classes like Psychiatry / Psychology and Radiology have lower accuracy, possibly due to factors such as fewer training samples or overlap with other specialties. The number of entries column also provides information about the samples' distribution and data imbalance. Departments that have more entries and misclassifications, like Neurology and Radiology, could need more data or better-quality data to increase the degree of accuracy. For the enhancement of the model, some of the steps include the introduction of domain knowledge, more discriminative features, hyperparameter optimization, and advanced data augmentation. Misclassified instances of medical classes might reveal further insights into prevalent issues and areas for targeted development and improvement. The number of BioBERT misclassifications based on the true medical specialty category indicates that Radiology had the most misclassifications (8), followed by Neurology (7) and Orthopedic (5). Cardiovascular/Pulmonary had three misclassifications. Hematology-Oncology, Psychiatry/Psychology, Pediatrics - Neonatal, and Urology each had two misclassifications. Obstetrics/Gynecology and Gastroenterology each had one misclassification. This illustrates areas where BioBERT's performance could be improved, particularly in specialties with significant misclassification rates, which include Radiology and Neurology, emphasizing the need for further data to increase model accuracy in these fields.

3.8.3 Result Comparison between Baseline and Advance model

Table 14. A comparison between baseline and advance model of medical specialty categories F1 scores

Medical specialty	BioBERT	ClinicalBERT	BERT	RoBERTa	Random Forest	Support Vector Machine	XGBoost	Logistic Regression
Cardiovascular / Pulmonary	0.94	0.93	0.94	0.94	0.67	0.69	0.72	0.75
ENT – Otolaryngology	1.00	0.97	0.97	0.97	0.96	0.95	0.97	0.97
Gastroenterology	0.96	0.98	0.86	0.98	0.83	0.87	0.87	0.89
Hematology – Oncology	0.91	0.89	0.88	0.86	0.88	0.91	0.88	0.90
Neurology	0.88	0.88	0.85	0.85	0.45	0.51	0.52	0.59
Obstetrics / Gynecology	0.91	0.92	0.90	0.91	0.89	0.89	0.90	0.91
Ophthalmology	1.00	0.97	0.97	0.97	0.99	0.99	0.99	0.99
Orthopedic	0.95	0.93	0.92	0.92	0.71	0.75	0.74	0.76
Pediatrics – Neonatal	0.85	0.74	0.81	0.89	0.88	0.89	0.88	0.91
Psychiatry / Psychology	0.80	0.80	0.80	0.86	0.99	0.99	0.99	0.99
Radiology	0.91	0.90	0.92	0.90	0.22	0.27	0.36	0.40
Urology	0.98	0.97	0.92	0.93	0.87	0.91	0.89	0.91
Weighted Average	**0.93**	**0.92**	**0.90**	**0.92**	**0.77**	**0.79**	**0.80**	**0.82**

Source: Own results.

Table 14 also presents the comparison of the F1 scores of the baseline and advanced models for the medical specialty categories. The model's performance was measured using three metrics: precision, recall, and the F1 score. The weighted average F1 score is a performance metric for models that considers both precision and recall. We discovered that the BioBERT, ClincalBERT, BERT, and RoBERTa models had weighted average F1 scores of 0.93, 0.92, 0.90, and 0.92. This is a considerable improvement over prior research utilizing different machine learning algorithms discussed in section 3.8.1. One such study used the random forest algorithm, support vector machine, XGBoost, and logistic regression, resulting in weighted average F1 scores of only 0.77, 0.79, 0.80, and 0.82, respectively. The outcomes are comparable to those of previous investigations. Overall, our results show that pre-trained BERT models outperform non-transformer-based techniques. The BERT models outperform typical ML approaches in the present investigation due to their superior accuracy, as evidenced by the higher F1 score.

3.9 Explainable AI Technique

In the context of a BioBERT-based classification model for the medical specialty classification problem, we chose SHAP to understand model decisions and improve transparency. Explain prediction: SHAP attempts to bring together some of the available methods for analyzing machine learning model predictions. It is based on the Shapley values from cooperative game theory, which provide an approach to fairly sharing the "payout" (the prediction) among the "players" (the feature) (Lundberg & Lee, 2017, p. 3). SHAP values in our BioBERT model assist in comprehending how every word or phrase in a medical transcription contributes to the predicted medical specialty. SHAP,

by providing an importance value to each feature, allows us to visualize and quantify the impact of individual words on the overall prediction.

3.9.1 Importance of SHAP for Enhancing Medical Text Classification Transparency

- Enhanced Transparency: SHAP improves the transparency of medical AI by explaining each feature's contribution to the model's predictions, available in global and local explanations. This clarity assists clinical confirmation and substantially enhances the patient's satisfaction, hence improving patient health (Goodwin et al., 2022, p. 3). This two-pronged (global and local) explanations approach helps the healthcare givers affirm the models, gain confidence, and enhance patient outcomes by explaining the features that the model considered to arrive at the final decision.
- Trust and Accountability: When machine learning models are applied in healthcare, it is paramount to ensure that the algorithms' forecasts are reliable for physicians. SHAP values in this regard serve to restore this trust by showing which parts of the text are most influential in the decision-making process and thus making the model's behavior more transparent and understandable.
- Error Analysis and Debugging: SHAP can help discover any errors or biases in the model by highlighting the essential features that influence a prediction. For example, if the model depends excessively on unnecessary or deceptive phrases, it can be a sign that the training data or the model itself requires adjustment (Feigl et al., 2022, p. 2).
- Improved Model Understanding: Understanding how a model arrives at a particular prediction is critical to the improvement of the model. SHAP can identify relationships and characteristics that might not be easily observable from the actual data or even from the model results of predictions.
- Communication with Stakeholders: SHAP's visual explanations allow easy communication with non-technical stakeholders such as clinicians, administrators, and regulatory bodies. It can assist in increasing the class cooperation and acceptability of models in the application areas of real life.

3.9.2 Integration of XAI with BERT Model

Combining the SHAP XAI approach with BioBERT for medical specialty categorization aims to provide transparent information about the model's decision-making, which should be highly appreciated in the delicate realm of healthcare. To classify medical transcription text data, BioBERT, a pre-trained model, is applied to biomedical text. In this work, after training on medical datasets, SHAP values are applied to explain and visualize the token or word's contribution to the model. This procedure helps explain why BioBERT predicted that particular result to bring about the decision that was made, and it combines the XAI and BERT models as depicted in Figure 29. Based on the BioBERT model and prediction functions, a sample from the test set of medical text data will be chosen and then fed into the SHAP Explainer. The explanation computes SHAP values that concern the given input and

provides the weighed lift of each token in the prediction of BioBERT. Visualizing these SHAP values is a key step because the decision-making process becomes easier to understand where the tokens are affecting the model's output, thus allowing for better comprehension of the model's reasoning and making the decision process more accessible.

Several obstacles must be overcome before XAI can be effectively integrated with BioBERT. BioBERT interpretation is complicated due to its multiple parameters and sophisticated token relationships, necessitating careful visualization and interpretation. To avoid mistakes, ensure that the tokenizer used for SHAP explanations corresponds to the one used during model training. In addition, computing SHAP values is computationally intensive because it involves many forward passes through the model. Some important issues are as follows: the main is how to handle sequences that are longer than BioBERT's token limit, and the other is how to increase the model's performance while keeping its interpretability. It is noted that data processing, which is performed when designing an SHAP explanation model, should be coherent to derive sensible results for medical practitioners. (Ngai & Rudzicz, 2022, p. 5). By overcoming these issues, XAI-BioBERT integration can improve confidence and transparency in medical specialty categorization tasks.

3.9.3 Validation of XAI Result and Robustness Testing

A crucial step to ensure the reliability and interpretability of the model predictions is the verification of the results obtained by the XAI, especially by SHAP. In other words, it means ensuring that the changes highlighted in the features affect the decisions of the model and align with the knowledge of the medical domain. One way of bolstering confidence in the model is by proceeding to perform a detailed analysis of the internal and external validity of the SHAP values across different samples. A robustness check helps to check that the performance of the model is not very sensitive to extreme values or does not deteriorate in some circumstances to increase confidence in the ability of the model to be used in real usage.

a) Validation of SHAP technique result: For the purpose of ensuring that the predictions provided by the BioBERT model are reliable and robust, it is necessary to verify the results of the SHAP technique. In order to show that the model is capable of successfully identifying significant words, maintaining consistency across samples, and generalizing well to a range of inputs, we will be examining SHAP values that have been obtained from medical specialty cases.

- Cross-Validation on SHAP Analysis: It is essential that the SHAP analysis be consistent throughout many folds to preserve its credibility. The outputs that are shown in Figure 32 are the calculated SHAP values for several medical specialty instances, including Cardiovascular/Pulmonary, Orthopedic and Urology cases. This indicates that the model has been assessed on an extensive range of different inputs.

Figure 32. Dataset samples with SHAP values across medical specialties (Cardiovascular/Pulmonary, Orthopedic, Urology) and color-coded word significance (red: positive influence, blue: negatively influence, white: neutral influence) and BioBERT model predictions

Source: Own results.

The model's prediction for the Cardiovascular/Pulmonary specialty is consistent with domain-specific terms like "ventricular," "mitral," "ejection fraction," and "doppler," suggesting accuracy. Similarly, for the Orthopedic specialty, the existence of phrases such as "carpal tunnel," "ligament," "tendon," and "incision" verifies the model's prediction, as these terms are associated with orthopedic treatments. In the discipline of urology, the significance of words such as "flank pain," "ureteral," "stone," and "shockwave" supports the model's prediction accuracy.

- SHAP Value Examination: The words with higher SHAP values, shown in red in Figure 32, are expected to be associated with the predicted medical specialization. The SHAP value evaluations demonstrate that the model exhibits a high level of accuracy and interpretability in three specific medical specialties: Cardiovascular/Pulmonary, Orthopedic, and Urology. The Cardiovascular/Pulmonary class has a base value of 0.0334385 and a SHAP value of 0.970981, indicating the model's strong confidence in its prediction. The model's high SHAP score in this domain is attributed to the recognition of important terms such as "ventricular," "mitral," "ejection fraction," and "doppler," which are necessary for cardiac functions and diagnostics. This highlights the model's precision in this field.

  The Orthopedic class has a base value of 0.165611 and a SHAP value of 0.96804, respectively, which reflect the high degree of confidence that the model has in its predictions. The model places an emphasis on technical terms such as "carpal tunnel," "ligament," "tendon," and "incision," all of which are crucial to orthopedic issues and surgical therapies. It is clear that the model is capable of identifying orthopedic diseases, as shown by the high SHAP value and the consistent use of relevant wording.

  In the Urology class, the model starts out with a baseline value of 0.0492639 and concludes with a SHAP value of 0.961569. The inclusion of key phrases in the field of urology, such as "flank pain," "ureteral," "stone," and "shockwave," demonstrates the model's ability to concentrate on pertinent aspects of the medical specialty. The SHAP values indicate that these terms have a significant impact on the model's predictions, hence strengthening the model's reliability and practicality.

  In a nutshell, the SHAP visualizations in all three specializations validate the model's predictions since they are relevant, consistent, and robust in explaining SHAP values. Every projected phrase aligns with the predicted terminology in the medical field for the specific profession, hence enhancing the overall dependability of the model.

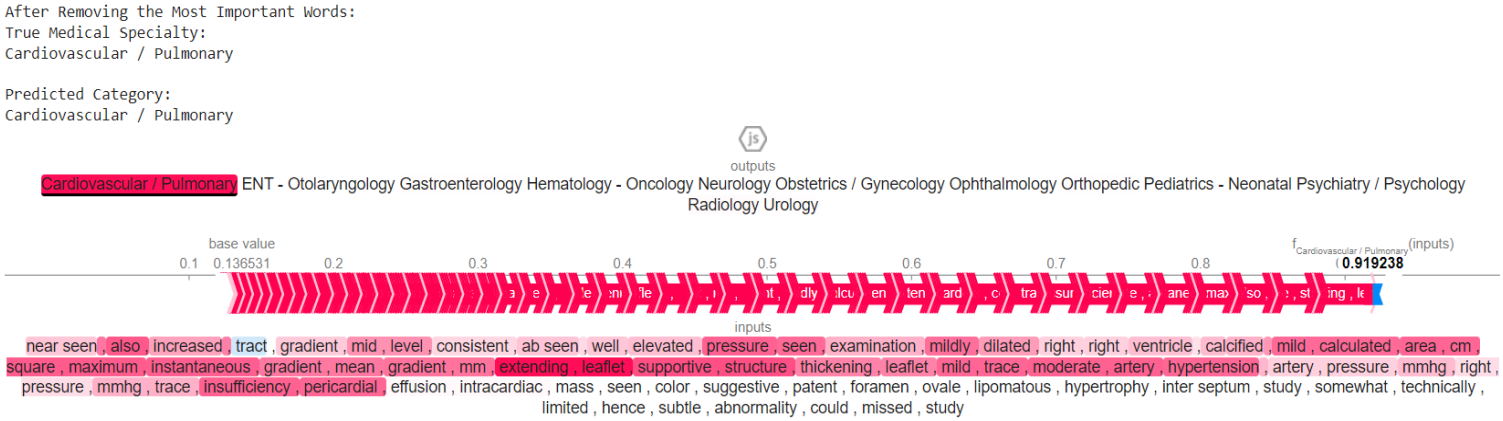
- SHAP Value Consistency Check: In order to evaluate the dependability of prediction models, it is crucial for SHAP values to exhibit consistency over several samples. Figure 32 demonstrates consistency by showcasing sentences that consistently obtained favorable SHAP ratings across many categories. The terms "ventricular," "echocardiogram," "aortic valve," and "pulmonary" are often used in Cardiovascular/Pulmonary projections, highlighting their importance in this medical

field. Similarly, in the field of orthopedics, the terms "carpal tunnel" and "ligament" are often recognized as significant components, confirming their ongoing value. The consistent use of terms such as "flank pain" and "ureteral" in urology predictions underscores their enduring significance in predicting the outcomes of urological conditions. The results suggest that SHAP values are successful in identifying and ranking important elements in many medical domains, hence enhancing trust in the model's prediction capabilities.

- Consistency Over Multiple Samples: Consistency over different samples is critical for guaranteeing that a model can generalize properly. The SHAP values reported show a consistent pattern across different medical fields. In Cardiovascular/Pulmonary situations, pertinent medical language is consistently used across instances, illustrating the model's ability to consistently detect and prioritize relevant elements. Similarly, in orthopedic contexts, relevant terms such as "carpal tunnel" and "ligament" are constantly highlighted across several inputs, demonstrating the model's capacity to recognize essential features within orthopedic data. Furthermore, in Urology, phrases such as "flank pain" and "ureteral" regularly emerge as key features across various urology-related inputs, demonstrating the model's capacity to consistently identify and weigh important characteristics in this medical specialty. This consistency in recognizing and leveraging key phrases across multiple samples improves the model's overall robustness and ability to give correct insights across a variety of medical contexts.

b) Robustness Testing of SHAP Technique: Evaluating the robustness of machine learning models is crucial for their practical implementation, since only resilient systems can be relied upon to function consistently. In this context, robustness is the term used to describe a model's ability to perform well even when faced with noisy or interrupted data. Model resilience is often evaluated by examining its accuracy on test data that includes noise. An often-used method involves generating test data with noise via the utilization of domain-specific input generation methods. These methods introduce modifications to pre-existing, high-quality seed data (Lambert et al., 2023, p. 159). Perturbation testing is a method used to evaluate the robustness of a model by eliminating the most crucial words identified by SHAP and then reassessing the model's predictions. This research is a comprehensive analysis conducted using the findings and SHAP map shown in Figure 33.

Figure 33. Perturbation testing of the selected sample by removing the most important words identified by SHAP

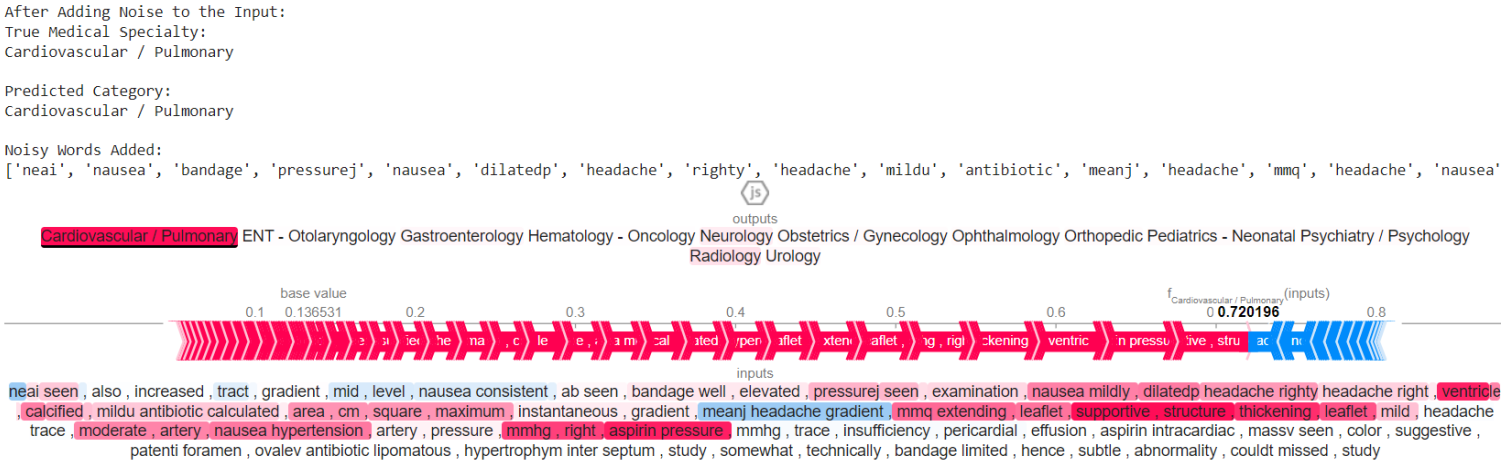


Source: Own result.

The objective was to assess the effect of removing crucial predictive terms on the performance of the model. Using the provided sample data, the model first made a prediction about the category "Cardiovascular / Pulmonary" with high confidence, relying heavily on terms such as "cardiovascular", "pulmonary", "doppler", "echocardiogram", "annular", "aortic", "root", "aortic", "valve", "atrial", "atrium", "calcification", "cavity", "ejection", "fraction", "mitral", "obliteration", "outflow", "regurgitation", "relaxation", "pattern", "stenosis", "systolic", "function", "tricuspid", "valve", "ventricular", "ventricular", "cavity", "wall", "motion", "pulmonary", "left", "ventricular", "cavity", "size", "wall", "thickness", "appear", "normal", "wall", "motion", "left", "ventricular", "systolic", "function", "appears", "hyperdynamic", "estimated", "ejection", "fraction", "near", "cavity" and "obliteration". After removing these important words, the model still correctly predicted "Cardiovascular / Pulmonary," but the confidence decreased from 0.970981 to 0.919238. This suggests that, while the model's predictions remained robust, the lack of important terms reduced its confidence level. After eliminating crucial words from the SHAP plot, terms such as "mildly," "dilated," "right," "ventricle," "extending," "leaflet," "artery," and "hypertension," (highlighted in red) still had a substantial influence on the model's performance.

Although confidence decreased, the model demonstrated its robustness by accurately predicting the category even in the absence of essential terms. Additional assessment included introducing noisy words into the same sample from which significant words had been eliminated. Noisy words such as "aspirin," "antibiotic," "bandage," "nausea," and "headache" were added, along with slight misspellings of existing words, such as "neai" (misspelled), "pressurerj" (misspelled), and "dilatedp" (misspelled). The model demonstrated a high level of confidence in its predictions, effectively managing irrelevant data and noise. This additional noise did not affect the words "mildly," "dilated," "ventricle," "artery," and "hypertension" (marked in red in Figure 34) from shaping some aspects of the model; confidence dropped from 0. 919238 to 0. 720196; however, texts colored blue shall have a negative impact on the prediction. Consequently, this robustness assessment illustrates the model's capacity to analyze the whole context and make sound predictions even if some information is absent or noisy.

Figure 34. Perturbation testing of a selected sample of the most important words removed, which are identified by SHAP, and adding noisy words



Source: Own result.

The perturbation test indicates that the model does not rely heavily on a small number of particular phrases. Instead, it has acquired an understanding of the overall context of medical texts and is able to maintain accuracy even when the inputs change. This demonstrates the model's ability to comprehend the language peculiar to the medical domain. The robustness of the model demonstrates its dependability and ability to function in real-world scenarios. This implies that it can accurately categorize medical specialties, even in the absence of important words or the inclusion of minor phrases. Consequently, it can identify the medical specialization by analyzing a combination of words and phrases, which is crucial for handling varied and complex medical transcriptions.

3.9.4 Error Analysis using Integration of BioBERT with XAI

Error analysis utilizing BioBERT and Explainable AI provides a transparent method for evaluating and improving model predictions in medical specialty classification. We can use SHAP values to discover significant elements impacting the model's judgments as well as regions where the algorithm misclassifies data. This technique improves our ability to fine-tune models as well as their contextual knowledge and accuracy (Ngai & Rudzicz, 2022, p. 3). The error analysis utilizing SHAP values exposes key information about why the BioBERT model misclassified the medical specialization. In this example of selected sample data, as shown in Figure 35, the correct label was "Radiology," yet the model predicted "Orthopedic." The SHAP illustration shows how words like "left," "digit," and "numbness" greatly influenced the model's prediction of "Orthopedic." These phrases are strongly connected with orthopedic problems, especially those concerning the musculoskeletal system. The model most likely concentrated on these terms while ignoring the larger context that would signal a requirement for radiological expertise. This mistake implies that, while BioBERT can efficiently identify essential terms connected to single specialties, it may struggle to recognize nuanced context when numerous specialists use the same terms.

Figure 35. Error analysis of a misclassified instance of a medical specialty sample using the integration of BioBERT with XAI

Source: Own results.

The main reason for this misclassification appears to stem from the model's focus on individual phrases as separate entities without considering the clinical context sufficiently when identifying them, especially in the case of "Radiology." This suggests a deficiency in the model's capacity to consider context in its training data, potentially due to its lack of training to distinguish between various specialized fields. The model could potentially improve by incorporating more training data and more elaborative definitions of the parameters for each profession. Furthermore, incorporating additional features from clinical scenarios and diagnosis, such as patient history, symptoms, and specific clinical outcomes, into feature engineering will enhance the model's performance. These enhancements have the potential to advance the model to a higher level of learning and reduce misclassification, hence improving its efficacy in treating various medical specialties. The SHAP values provide a direct assessment of the model predictions made using the BioBERT and XAI combination. It emphasizes the need for having a thorough understanding of the context and using better training data to reduce misclassification. An examination of the SHAP values allows for the identification of critical areas for improving the model and refining the data, resulting in improved performance and accuracy in the classification of medical specialties.

3.10 Validity and Reliability of Research Methodology

The study's validity and reliability are essential for generating trustworthy and dependable outcomes of the research process. We adhere to the subsequent procedures to ascertain the validity and reliability of the investigation's methodology. These procedures enhance the effectiveness and accuracy of research gathering practices, ensuring that the study is grounded in solid and trustworthy facts. Additionally, these procedures aid in preserving the authentic and precise nature of the investigation's process and results.

a) Data Collection Validity

- Source Selection: The dataset was obtained from mtsamples.com, a comprehensive repository of medical transcriptions. This decision ensured a diverse range of specializations and transcription varieties, hence enhancing the dataset's relevance.
- Comparative Advantage: The chosen dataset offered a wider and more diverse range of transcriptions compared to other limited datasets like TCM, Hallmarks, and AIM, enabling a more comprehensive evaluation.

b) Dataset Overview and EDA Reliability

- Dataset Composition: The dataset contained a total of 4999 records and six columns, which provides a substantial sample size for analysis. The focus on important columns like medical_specialty, transcription, and keywords ensured that the analysis considered relevant and valuable data points.
- Exploratory Data Analysis: A combination of graphical and non-graphical methods led to the discovery of significant patterns, resulting in a comprehensive and valuable analysis. We achieved a comprehensive and reliable understanding of data distribution and important features by utilizing various visualization techniques like bar plots and word clouds.

c) Data Preprocessing Steps for Reliability

- Handling Missing Values: In the "keywords" column, we replaced missing values with empty strings and removed entries with missing transcriptions column data. This ensured that the dataset was fully comprehensive and prepared for analysis.
- Text Normalization: To maintain data consistency and reduce noise, we transformed the text to lowercase and removed all punctuation and numbers.
- Tokenization: Tokenization is the process of splitting down the text into smaller units known as tokens. This procedure helps to make the data consistent and easier to handle during processing and analysis.
- Stop Words Removal and Lemmatization: Eliminating frequent, unimportant words and reducing words to their basic forms enhanced the significance of the text data, making it more relevant for analysis.

d) Ensuring Validity in Data Analysis

- Category Filtering and Adjustment: Categories with less than fifty samples were eliminated to focus the dataset on relevant and well-represented specialty.
- Label Encoding: The conversion of medical specialties into numerical values facilitated a standardized modeling method, enhancing the analysis accuracy and reliability.

e) Model Selection Validity

- Baseline and Advanced Models: Various machine learning models, including Random Forest, SVM, XGBoost, and Logistic Regression, as well as advanced deep learning models like BERT, BioBERT, Clinical BERT, and RoBERTa, were employed. This approach ensured a comprehensive method for selecting models, incorporating the strengths of various techniques.
- Error Analysis and Misclassification Handling: Errors have been detected and examined using confusion matrices, feature importance analysis, and SHAP values. This thorough investigation helped to gain a better understanding of and find solutions for model limitations.

f) Explainable AI for Reliability

- SHAP for Model Transparency: Utilizing SHAP values, we were able to interpret the contributions of individual words or phrases to the model's predictions, resulting in increased transparency and trust.
- Validation and Robustness Testing: The reliability and consistency of the SHAP analysis were confirmed through cross-validation. The model's resilience was demonstrated through perturbation testing, which assessed its performance when important terms were missing or noise was introduced.

g) Ensuring Trustworthiness of Results

- Data Quality Assurance: Cleaning and preprocessing the dataset made sure it was error-free and correctly prepared, ensuring the analysis used reliable data.
- Model Validation and Cross-Validation: Employing cross-validation techniques ensured that the model's performance remained consistent and avoided overfitting to a particular subset of data. Error analysis and correcting processes, such as targeted retraining and data augmentation, enhanced the model's accuracy and dependability.
- Explainable AI Integration: The incorporation of SHAP enabled model predictions to be assessed, providing transparent insights into the model's decision-making process. This ensured that the outputs were easily understandable and reliable.
- Robustness Testing: Robustness testing entails carrying out perturbation testing in a variety of scenarios, such as removing important words or introducing noise. This testing helps validate the model's stability and dependability in real-world applications.
- Consistency and Reproducibility: The use of standardized processes for data preparation, model training, and assessment ensured consistency and reproducibility, preserving uniformity and the ability to replicate the research approach. Providing a detailed account of every stage in the thesis research procedure enhanced clarity and the ability to replicate the results.

By meticulously following these stages, the research methodology ensured the validity and reliability of the results, thereby establishing the credibility and trustworthiness of the findings. This comprehensive methodology enhances the integrity of the research process and guarantees that the conclusions obtained are based on solid and trustworthy evidence.

3.11 Ethical Considerations

Some of the identifiable ethical issues in medical data research include those that relate to medical transcriptions from mtsamples.com, which are essential in ensuring that patient's identities are protected, and the data shared is accurate and impartial. This implies that if proper precautions are taken with the sensitive material as well as being truthful in the data used, the use of XAI methods benefits the accountability to the model predictions (Williamson & Prybutok, 2024, p. 7). To ensure ethical research that upholds patient rights and remains objective, it is essential to strike a balance among these elements.

a) Informed Consent and Data Usage

The use of medical transcription data raises important ethical considerations regarding the confidentiality and privacy of patients. It is crucial to choose sources such as mtsamples.com, which anonymizes and aggregates data, to address these concerns. It is essential to prioritize rigorous anonymization and strict adherence to medical data handling guidelines in order to safeguard patient information, as highlighted by (Rodriguez et al., 2022, p. 1). If mtsamples.com adheres to legal and ethical standards, its dataset can be used in an ethical manner.

b) Data Transparency and Usage Context

Full disclosure of the data's origin, any biases, and use constraints is crucial in research. It is essential to provide a comprehensive explanation of the dataset and recognize its constraints in order to maintain the integrity of the study. By offering a comprehensive description of the investigation's objectives, such as the analysis of medical transcriptions for the aim of categorization, and explicitly acknowledging any constraints in the dataset, it fosters confidence in the results by addressing any biases.

c) Bias and Fairness

To achieve fairness in medical research, it is necessary to address biases in the collection of datasets and the predictions made by models. Having a wide range of diverse and representative datasets in different medical specialties is essential. Additionally, it is vital to extensively evaluate the performance of models across various demographic groups and analyze errors using SHAP values. These measurements aid in detecting and resolving prediction biases, with the objective of building a more fair and accurate system in medical applications.

d) Model Interpretability and Accountability

For crucial medical applications, black box models raise serious ethical concerns due to their tendency to make decision-making more difficult by making it challenging to explain and understand their predictions. If we want to solve this challenge, we need to understand the model using

explainable AI approaches like SHAP values. This approach improves interpretability, which makes it easier for the stakeholders, such as physicians, to trust the model results. Furthermore, error analysis, particularly in the context of medical specialties classification, aids in pinpointing misunderstandings and enhancing the efficiency of the model. Despite the time-consuming nature of this rigorous examination, it enhances the model's accountability and reliability, thereby boosting the system's overall trust.

e) Research Integrity and Reproducibility

It is also important to have clear records of the data collection, data preprocessing steps, model training process, and model evaluation results. Cross-validation and robustness testing introduce additional tests for the model's performance in different contexts, thereby increasing research credibility. In total, these practices ensure that the information obtained from research is accurate and reliable.

The research approach aims to uphold ethical practices in medical data research by acknowledging these ethical issues and implementing appropriate measures to address them. This method safeguards patients' anonymity and establishes standards for fair and efficient utilization of the collected data in healthcare AI applications.

3.12 Limitations

a) Sample Size and Representativeness: The analysis may have missed the diversity of medical transcriptions in all fields, despite only having 2,324 records in the study. In the case of underrepresented specialties in particular, this restriction limits the model's ability to generalize to new data. As a result, the model may perform less accurately and significantly less straightforwardly in these domains.

b) Data Biases: Issues related to data representation, for instance, an imbalance between "surgery" and "hospice - palliative care," could affect the results of models. When it comes to overfitting, certain specialties experience classification at the expense of less frequently sampled specialties, while oversampled specialties experience underperformance. This dampens the accuracy and dependability of the model, especially when applied to real-life situations.

c) Data Quality and Anonymization: Concerns about privacy and ethics arise when anonymization procedures lack confirmation. Insufficient anonymization renders the data unsuitable for use in clinical care settings, thereby rendering the study unethical. This strongly impacts the effectiveness and relevance of the obtained data.

d) Contextual Understanding and Domain Specificity: Contextual factors of medical language can pose challenges in terms of classification accuracy during the preprocessing step and model training. In order to overcome these difficulties, new approaches in preprocessing have to be developed, and

changes need to be made that are relevant to the individual domain. By doing so, we ensure that the model has some ability to handle medical transcription compounds.

e) Tokenization and Text Normalization: The process of overnormalization, which involves converting all text to lowercase and removing punctuation or numerals, can aid in removing some crucial information from the medical transcriptions. Such data loss impacts the evaluation and classification process of the model. Therefore, we should implement a well-balanced normalization technique to preserve the original data of the medical condition.

f) Ethical and Privacy Considerations: Making assumptions about anonymization can easily overlook some ethical issues, such as informed consent and patient privacy. These breaches may threaten the credibility and acceptability of the findings of research. Maintaining integrity and trust requires a commitment to the highest ethical standards and ensuring transparency and observability in all data practices.

These major limitations make it clear how important sample size, data bias, data anonymization methods and procedures, knowledge of the context, and ethical concerns are when creating and using medical transcription models. While the overall design of this study is sound and reasonably inclusive, it is important to declare potential biases. These limitations create opportunities for improvement and optimization; thus, future rounds of the study will be able to respond to these considerations, increase the credibility of results, and provide more reliable and ethical approaches to medical text classification.

3.13 Summary

This study's research technique strictly aimed to measure and categorize medical text transcriptions in accordance with its objectives. We collected data from mtsamples.com, compiling a total of 4999 medical transcription samples, which appeared to be broader and more extensive than other datasets from other websites. We used some of the basic, additional, and focus columns of investigations, including medical specialty, transcription, and keywords. This study used both graphical and non-graphical means to understand the structure of datasets and other patterns of relevance, and the tools included bar graphs and Word clouds, among others. Some of the steps performed were missing value treatment, text standardization, tokenization, filtering, and lemmatization of words, which improved data quality and reduced noise. We reduced the number of records to 2,324, requiring only a few general categories for the model training to guarantee robust representation.

We selected a variety of classic machine learning models, including Random Forest, SVM, XGBoost, and Logistic Regression, as well as advanced deep learning models like BERT, BioBERT, Clinical BERT, and RoBERTa, due to their ability to enhance classification accuracy. We used SHAP values, a component of explainable AI, to interpret the results for the learning model's predictability and identification of misidentified classes. Pre-processing on the dataset, cross-validation, and fine-

tuning enhanced the validity and dependence of the feature selection on the dataset. More specifically, we also discussed other topics such as data privacy, data transparency, and biases to build the ethical groundwork needed when dealing with patients' sensitive information. It could be so, as this approach was logical and sequential, with the tasks unambiguous and well thought out; the outcome of the analysis was complete and rigorous; and the model's predictions were authentic and dependable in concurrence with the research goals and objectives.

Chapter 4: Research Findings and Interpretation

4.1 Introduction

The chapter on Research Findings and Interpretation offers a detailed evaluation of the research study and focuses on enhancing the reliability of texts about medical matters by assessing the BERT models and XAI approaches to making the decisions. This chapter is dedicated to a detailed consideration of the outcomes obtained from advanced NLP techniques for further analysis of a dataset of medical transcription samples regarding data preprocessing, EDA, model performance evaluation, and the use of XAI. The findings demonstrate the superior performance of transformer-based models such as BERT over typical machine learning algorithms in handling complicated medical data, while underscoring the crucial role of XAI in increasing model interpretability and trustworthiness. Using approaches such as SHAP analysis, the study shows how clear visual explanations of model predictions can improve clinical decision-making and reveal potential biases. The chapter not only demonstrates the effectiveness of advanced NLP models in healthcare informatics, but it also emphasizes the significance of incorporating domain-specific knowledge and robust preprocessing to assure the development of clinically appropriate and adaptable AI systems. These insights make a substantial contribution to the evolution of AI-driven healthcare applications, enabling better patient care, medical research, and healthcare management through actionable insights drawn from unstructured medical data.

4.2 Research Findings

4.2.1 Dataset Characteristics and Preprocessing

The data used for this study is the medical transcription samples, which were extracted from mtsamples.com and contained 4999 samples, organized into six columns: Their columns include ‘Unnamed: 0,’ ‘description,’ ‘medical_specialty,’ ‘sample_name,’ ‘transcription,’ and ‘keywords’. Thus, in this research, these specific columns were excluded from analysis and kept only as ‘medical_specialty,’ ‘transcription,’ and ‘keywords.’ Moreover, ‘medical_specialty’ and ‘transcription’ were prioritized in the following analysis. When examined, the dataset had a total of 4999 records, and we discovered that some records were missing values in the 'keywords' and 'transcription' fields. More specifically, we managed 1068 blank entries under 'keywords' and 33 blank entries under 'transcription' by eliminating irrelevant rows in the 'transcription' column and replacing empty 'keywords' values with empty strings. After cleaning the dataset, we reduced it to 4966 records and three columns.

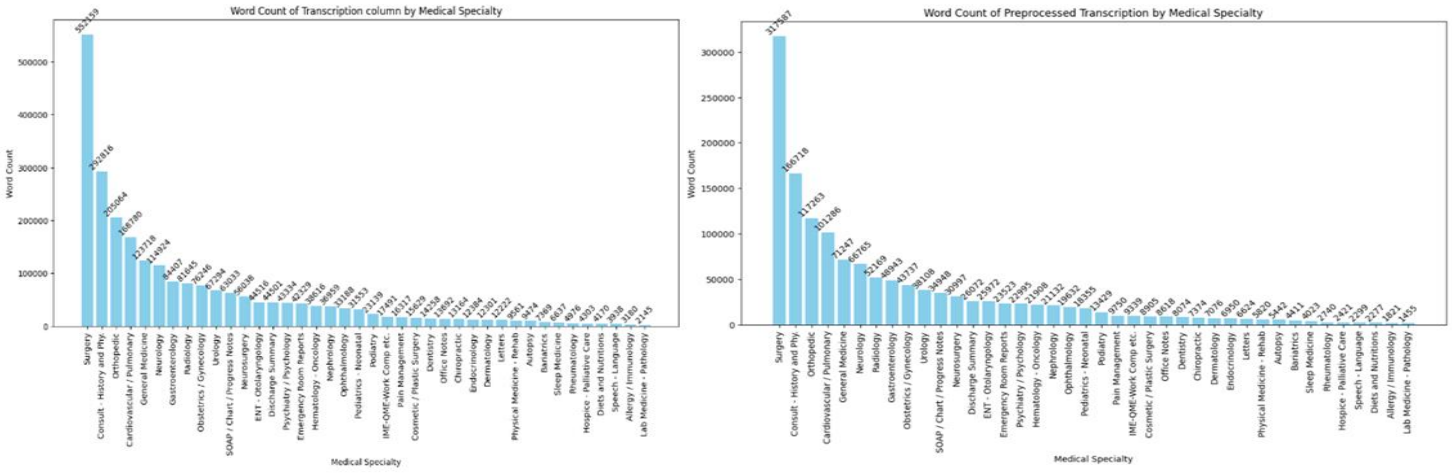
EDA revealed important insights into the distribution of medical specializations, the number of words related to each medical specialty, and the analysis of keywords. Surgery has emerged as the dominant medical specialty, with the highest number of terms in transcriptions. The emphasis is on the fluctuating amount of content and the use of keywords in different specialties. The EDA not only identified 'Surgery' as the medical specialty that appeared most frequently, but also provided

substantial insights into other specialties. Specialties such as 'Cardiovascular/Pulmonary' and 'Orthopedic' have large word counts, which suggests that they need detailed documentation and include complicated clinical cases. Specialized fields such as 'Hospice - Palliative Care' have reduced word counts, suggesting the presence of more streamlined and targeted documentation procedures within these sectors.

Analysis of keywords uncovered clear patterns within various medical specialties. Although 'Surgery' was the most frequently mentioned topic, 'Radiology' and 'Neurology' had distinct keyword sets that reflected their specific diagnostic focuses. These findings highlight the variation in content volume and thematic focus among different specialties, which is important for comprehending the intricacies of medical text data and providing guidance for future studies or model development in healthcare applications. Properly preparing textual data is absolutely essential in order to convert raw medical transcription data into a structured format that can be effectively analyzed and used for machine learning purposes. Preprocessing methods, like text normalization and noise reduction, have proven to be quite useful in uncovering the grammatical and semantic nuances of medical language.

The application of lemmatization and the elimination of unnecessary words significantly improved the dataset. This process enhanced the dataset by highlighting important content phrases and reducing linguistic inconsistencies. In addition, it provided valuable insights into the intricacies of medical transcriptions, which is a significant advantage. Figure 36 demonstrates that 'Surgery' has the largest word count before and after preprocessing, according to the transcription column. But there are fewer words overall, down from 552,159 to 317,587. A similar reduction in word count occurs in the 'Consult - History and Phy.' section, which goes from around 292,816 words to 166,718 words. The fact that the relative distribution has not changed despite these reductions demonstrates how well the preprocessing has removed noise without compromising the data's quality for use in further research.

Figure 36. Comparison bar graph of word count of transcription column before (left) and after (right) data preprocessing

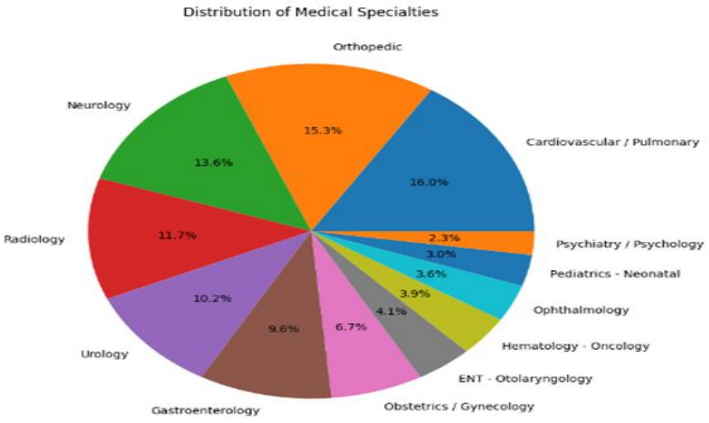


Source: Own result.

These findings are crucial for developing advanced natural language processing models capable of analyzing and categorizing medical material. This will greatly assist in healthcare administration, research, and clinical decision-making. We conducted thorough preprocessing on the dataset to ensure its accuracy and integrity, which sets the stage for advanced analysis. The practical insights uncovered by the investigation are driving new developments in healthcare informatics. Experts in the field carefully refined the dataset, enhancing its structure and ensuring its relevance to clinical practice. In order to accomplish this, we organized similar areas of expertise into broader categories, ensuring a more precise depiction of medical specialties. After refining the data, we had a total of 2,324 records and 29 unique specializations. This facilitated more equitable allocation and improved the resilience of our categorization model. Cardiovascular/Pulmonary had the highest count, while Hospice-Palliative Care had the lowest. Gaining a deeper understanding of the frequency of various specializations enhances the model's performance and ensures clinically meaningful insights.

The dataset underwent meticulous categorization to exclude medical specialties with a sample size of less than 50, hence guaranteeing precise evaluation and model training. Upon implementing all the filters, the dataset had a grand total of 2,324 records. The data were organized into two columns: "medical specialty" and "preprocessed transcription." The dataset includes a wide variety of medical disciplines. The collection included a wide range of fields, such as Cardiovascular and Pulmonary, Orthopedic, Neurological, Radiological, Urological, Gastroenterological, Obstetrics and Gynecology, ENT-Otolaryngology, Hematology-Oncology, Opthalmology, Pediatrics-Neonatal, and Psychiatry/Psychology. Figure 37 presents a visually appealing representation of the dataset's composition, specifically a pie chart illustrating the distribution of various medical disciplines. After data cleaning, the number of words in the transcription column dropped significantly from 2,407,470 to 629,262. Lemmatization, together with the removal of punctuation and stopwords, resulted in a decrease of 73.86%. The improved balance and representation in the dataset generated by these methods makes it more efficient and reliable, which further queries and classification models may benefit from.

Figure 37. Pie chart showing the distribution of more than 50 samples across various 12 different medical specialties



Source: Own result.

The dataset has undergone rigorous preprocessing and modification to enhance its suitability for use in machine learning applications. The dataset has a total of 2,324 elements, which are arranged in two columns. The preceding phase guaranteed an accurate depiction by using the knowledge of the medical domain to enhance the categorizations of medical disciplines. To ensure compatibility with machine learning algorithms that require numerical input, the category classifications were transformed into numerical values, resulting in the addition of a third column. The transcription data was processed to facilitate feeding into text vectorization tools. It was cleaned, normalized, and transformed from a nested list into a single list. Numerical features converted from text during the translation process were used in training the model. After adding the preprocessed transcription, medical specialty, and encoded target columns, machine learning activities on the dataset are now feasible. This will allow precise medical text categorization and prediction.

4.2.2 Model Performance

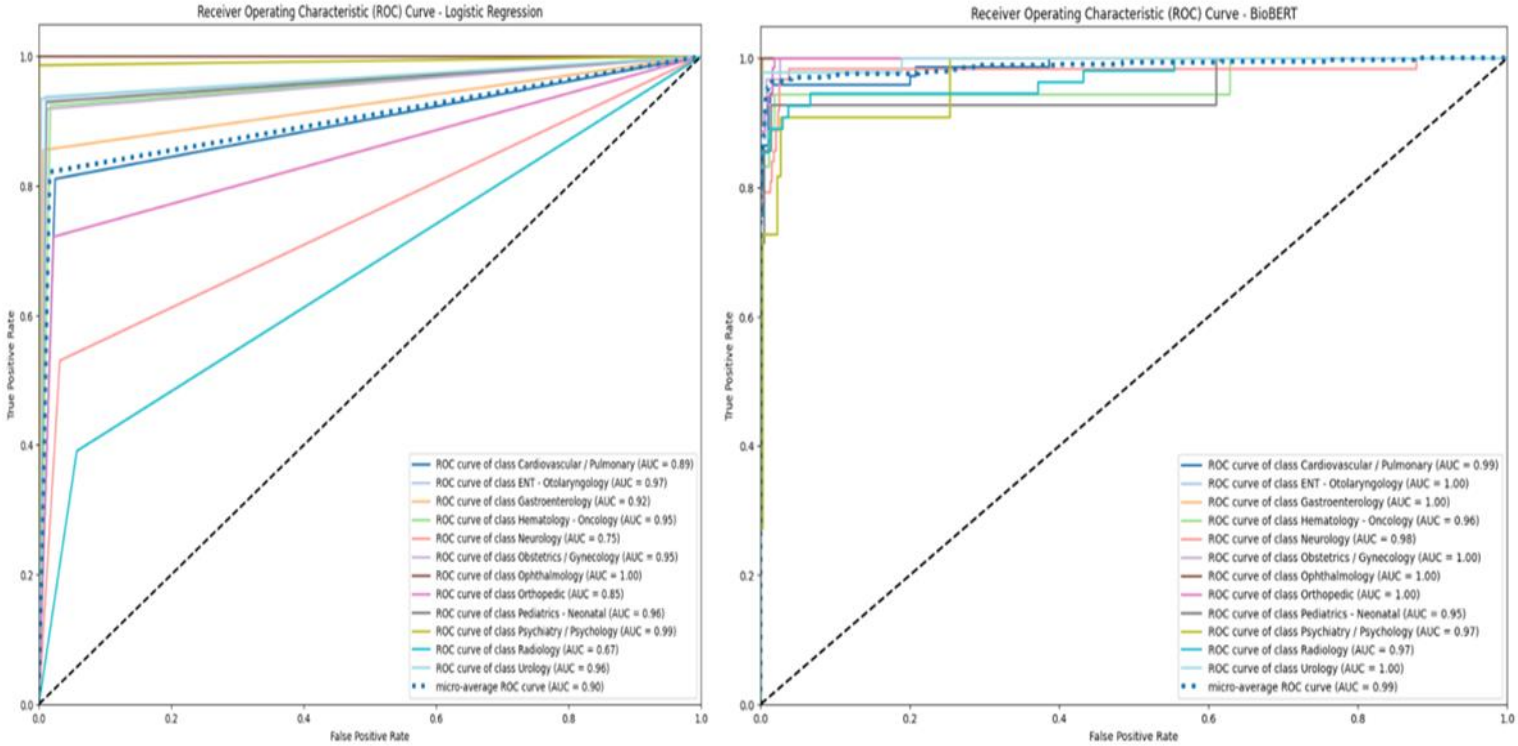
a) Effectiveness of Traditional ML Algorithms: Bag-of-words, TF-IDF, and bag-of-n-grams were among the many text representations used to assess classic ML algorithms like Random Forest, SVM, XGBoost, and Logistic Regression. Among them, XGBoost and Logistic Regression exceeded with an F1 score of 0.82 when using TF-IDF, therefore proving their resilience in handling tabular and structured data. With an F1 score of 0.80 with TF-IDF, SVM also performed well and shows excellent management of high-dimensional spaces. These results highlight the need for sophisticated text representation methods such as TF-IDF in enhancing the performance of conventional machine learning systems.

b) Impact of Text Representation Techniques: With TF-IDF appearing as the most effective approach among the conventional classifiers, routinely obtaining high F1 scores, text representation schemes had a significant effect on model performance. Less successful were bag-of-words and bag-of-n-grams, which underline the need for obtaining word relevance and context in text categorization applications. The advantage of TF-IDF is that models learn by knowing not only the presence of words but also their frequency and significance across texts, therefore improving their chances to find complex patterns in the data. This emphasizes the significance of using advanced text representation methods to get optimal outcomes in classification issues.

c) Superior Performance of Transformer Models: Transformer models (BERT, BioBERT, ClinicalBERT, and RoBERTa) outperformed all other models; BioBERT has a macro average accuracy of 0.94, recall of 0.91, F1 score of 0.92, and accuracy of 0.93. This exceptional performance shows that domain-specific transformer models, including BioBERT, may greatly enhance results in specialized domains such as biomedical text processing. Higher predictive accuracy results from BioBERT's specialized design and pre-training on domain-specific corpora, allowing it to capture nuances and complexity unique exclusively in biomedical literature. This emphasizes the possibilities of using domain-specific information for model development to achieve state-of-the-art performance in specialized uses.

d) Comparison of Transformer Models with Baseline Models: Plotting the receiver operating characteristic (ROC) curves and computing the area under the curve (AUC) can help one to evaluate the performance of baseline and advanced models, as seen in Figure 38. The ROC curves and AUC values show how much advanced transformer models outperform conventional machine learning models. For example, logistic regression revealed a range of AUC values (from 0.67 for radiology to 1.00 for ophthalmology) across classes. Similarly, a baseline model was generated. The sophisticated transformer model BioBERT showed outstanding comprehension of complicated patterns and connections in textual data, with most AUC values at or around 1.00. It got almost flawless AUC scores in all classes. This performance gap emphasizes how well sophisticated models can pick up and understand subtle language traits, hence producing much higher expected accuracy and reliability in specialized fields. These results show how transformational transformer-based models such as BioBERT may be in improving classification tasks as compared to conventional machine learning methods.

Figure 38. Comparison of ROC curves and AUC scores between traditional machine learning (linear regression) and advanced transformer models (BioBERT)



Source: Own result.

e) Data Handling and Model Training: Especially for minority classes, the usage of SMOTE was vital in correcting class imbalance by creating synthetic samples, hence improving model accuracy and robustness. CUDA-enabled GPUs significantly speed up model training, hence enabling more thorough investigation and improvement. These techniques highlight the requirement of advanced

handling of data and computational tools in modern machine learning systems, hence generating scalable and more effective models.

f) Evaluation and Cross-Validation: The consistent performance of BioBERT across data subsets was demonstrated by the 5-fold cross-validation, which further solidified its reliability for real-world applications. Still, the confusion matrix revealed issues with closely related classes that suggested data augmentation and improved training techniques were needed to reduce discrimination. These tests highlight the need for thorough validation and error analysis in improving the generalizability and accuracy of models, which are fundamental for their efficient application.

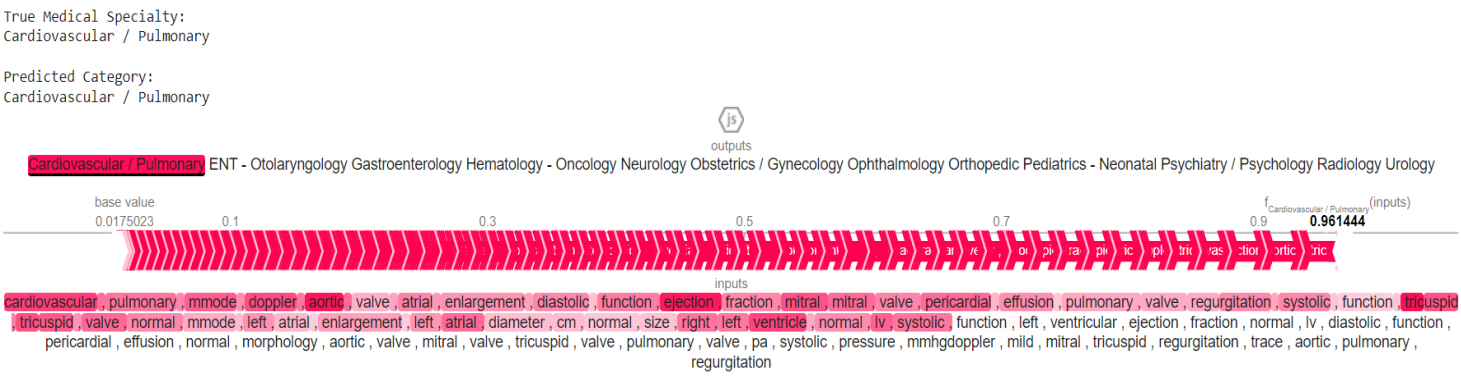
It is evident from the findings that more advanced deep learning models, especially transformer-based models like BioBERT, excel in comparison to typical ML algorithms on specific classification tasks. Domain-specific models have proven to be highly effective in various fields, offering valuable insights into their bright future. A comprehensive study of misclassifications helps one to get important insights to improve the performance of the model. By means of this study, one may find chances to strengthen the data and create more successful training plans. Studies show that when managing challenging and sophisticated textual input, advanced models show more degrees of accuracy and robustness.

4.2.3 Explainable AI Insights

a) Enhanced Transparency

- Clear Insights into Feature Importance: SHAP values provide comprehensive data on which features words or phrases drive the predictions of the algorithm. In the medical field especially, transparency is essential, as clinical validation and confidence depend on knowing the reasoning behind a categorization choice.
- Visualization of Contributions: Visual explanations explain the significance of each word to the model's choice, making the predictions easier to comprehend and trust. Figure 39 demonstrates the SHAP visualization of one sample transcription data, displaying the contribution of every word.

Figure 39. SHAP visualization of one sample transcription data showing the contribution of each word



Source: Own result.

b) Trust and Accountability

- Building Trust with Medical Professionals: SHAP values promote trust by highlighting the most influential elements of the text during the decision-making process. This is critical for ensuring that ML models in healthcare are credible and trustworthy for medical practitioners.
- Accountability in Predictions: Visualizing and quantifying the influence of individual words helps stakeholders understand how certain inputs produce particular results, therefore fostering responsibility.

c) Error Analysis and Debugging

- Identifying Potential Errors or Biases: SHAP is useful for finding biases or mistakes in the model. For instance, an excessive use of unrelated phrases by the model could suggest a compromise in either the training data or the model itself.
- Improving Model Quality: By highlighting critical aspects that impact the predicted result, SHAP debugs and enhances the model to achieve a more accurate and high-quality prediction.

d) Improved Model Understanding

- Revealing Patterns and Insights: SHAP values offer patterns and insights that raw data or model outputs alone may not reveal, allowing for a more in-depth knowledge of how the model produces predictions.
- Feature Relevance: The importance of features highlighted by SHAP values could assist in guiding future model development and data collection approaches.

e) Communication with Stakeholders

- Ease of Communication: Figure 39 presents the visual explanations offered by SHAP, which have proven effective in communicating information to individuals who may not have technical expertise, such as clinicians, administrators, and regulatory authorities. This enhances collaboration and acceptance of the model.
- Stakeholder Engagement: Enhanced transparency and interpretability can aid in securing stakeholder support and trust in the use of AI models in real-world healthcare settings.

f) Consistency and Robustness

- Consistent SHAP Values Across Samples: The consistency of SHAP values across samples, as illustrated in Figure 31 above, illustrates the model's trustworthiness. Consistently high SHAP scores in several categories indicate the model's strong feature detection.
- Robustness Testing: The model demonstrated its robustness by maintaining high prediction confidence levels despite phrase removal or word additions with noise. This exemplifies the model's versatility in handling intricate medical transcriptions.

g) Validation and Generalization

- Cross-Validation of SHAP Analysis: The accuracy of the model's predictions is guaranteed by regularly performing SHAP analysis over numerous folds.

- Generalization Across Diverse Inputs: The consistent distribution of SHAP values across medical specialties suggests that the model generalizes well to diverse inputs, increasing confidence in its implementation.

h) Error Analysis Using XAI

- Identification of Misclassification Causes: SHAP values aid in determining why the model misclassified instances, such as emphasizing individual phrases without considering the larger clinical context.
- Improvement Areas: SHAP-based error analysis findings emphasize the need for more training data or more advanced feature engineering to reduce misclassification and improve the model's contextual knowledge.

The combination of XAI approaches such as SHAP and BioBERT models provides considerable insights into model predictions, increasing transparency, trust, and responsibility. This method not only improves model comprehension and stakeholder communication, but it also aids in robust error analysis and debugging. As a result, using SHAP in medical text categorization assures that models are not only accurate but also interpretable and dependable, which is critical for real-world use in healthcare.

4.3 Interpretation of Results

4.3.1 Dataset Insights and Implications

Analysis and preprocessing of the medical transcription dataset from mtsamples.com produced some noteworthy findings with important implications for medicine and the use of NLP in healthcare.

a) Distribution of Medical Specialties

The original data collection included 4999 records from many medical fields. Preprocessing and domain-specific knowledge of the dataset reduced the dataset to 2324 records, each representing one of 29 distinct medical specialties. Reducing and refining works by removing specialties with fewer than 50 samples, resulting in a more balanced and focused dataset. Cardiovascular/Pulmonary, Orthopedic, Neurology, Radiology, and Urology are the disciplines most often represented. Out of all these, Cardiovascular/Pulmonary has the most records 371. Robust and adaptable NLP models depend on a balanced representation.

b) Word Count and Content Volume:

Throughout the preprocessing process, we reduced the number of words in the transcriptions from 2,407,470 to 629,262. This included eliminating stop words and punctuation, as well as applying lemmatization. This decrease (73.86%) demonstrates the efficacy of preprocessing in reducing background noise and focusing on key content. The research showed that fields requiring more detailed documentation, such as Orthopedics, Cardiovascular/Pulmonary, and Surgery, had greater word

counts. Hospice - Palliative Care, on the other hand, utilizes fewer words overall, which is indicative of their more simplified documentation criteria.

c) Keyword Analysis and Thematic Emphasis:

A keyword analysis revealed different patterns of use in many fields. Although surgery had the most phrases, other fields, including radiology and neurology, have unique sets of keywords reflecting their specific lexicon and diagnostic emphasis. This diversity in word use emphasizes the need to build NLP models that include the specific linguistic features of every medical discipline.

d) Data Quality and Preprocessing:

The rigorous preprocessing procedure, which included normalization, tokenization, and lemmatization, enhanced the dataset's quality and consistency. This preprocessing not only improved textual representation but also revealed the intrinsic complexity and nuances of medical transcriptions. A comparison of word counts before and after preprocessing revealed a significant reduction in noise while maintaining data integrity, which is essential for accurate and trustworthy analysis.

e) Clinical Relevance and Model Development:

Using domain knowledge to augment and filter the dataset significantly improved its structural and clinical relevance. Eliminating wide categories and combining related disciplines produced a more concentrated and intelligible dataset. This fine-tuning is important for the accuracy of subsequent NLP operations; it ensures that the dataset accurately captures the intended medical specialties. The balanced distribution of medical disciplines helps to build effective categorization models, which ultimately help to improve healthcare outcomes.

f) Transformation for Machine Learning:

The final dataset, which includes preprocessed transcriptions, medical specialties, and encoded target labels, is ideally suited for ML applications. The conversion of text data into numerical features enables excellent classification and prediction using medical text. This organized approach to data preparation not only enhances model performance but also assures that the insights produced are clinically relevant and useful.

g) Implications for Healthcare Informatics:

The findings from this dataset hold significant value for healthcare informatics. We can use the processed data to develop advanced NLP models that effectively categorize and assess medical language. These models can provide valuable insights from unstructured medical data, assisting with clinical decision-making, medical research, and healthcare management. In addition, examining the differences in the amount of content and the use of specific words in different specialties could assist in creating more customized and efficient NLP solutions in the healthcare field.

The findings from the investigation and preprocessing of mtsamples.com's medical transcription dataset are consistent with various themes and conclusions in the existing literature on dataset quality and its impact on AI model outcomes. According to Guo et al., (2016, p. 824), a balanced dataset is critical for creating robust and generalizable models capable of handling multi-label text classification while reducing biases and enhancing performance across multiple medical specialties. Supported by Ong et al. (2010, p. 2), who discovered that efficient preprocessing increases the performance of automated classifiers for clinical event reports, the significant reduction in word count after preprocessing shows the elimination of noise and unnecessary information. The better, more compact dataset helps models to concentrate on important traits, thereby improving prediction capacity and overall accuracy. Moreover, the variation in keyword use across specialties emphasizes the need for domain-specific models, a result that is compatible with Shao et al. (2018, p. 2877), who underlined the need for word embeddings in capturing specialized medical terminology. Qing et al. (2019, p. 8) emphasize the importance of thorough preprocessing, including normalizing, tokenizing, and lemmatizing, in improving data quality and consistency, resulting in greatly improved model performance. Changing the dataset assures appropriate representation of medical specialties by using domain knowledge, which is in accordance with Chai et al. (2013, p. 981), who gave clinical relevance high priority in data development. Devlin et al. (2019, p. 4173) research on BERT and other pre-trained language models demonstrates that this systematic approach to data preparation defines effective categorization and prediction. Healthcare informatics largely relies on high-quality input representations, as modern performance in NLP activities is dependent on them. Emphasizing the need for high-quality datasets in promoting innovation and improving patient outcomes, Lee et al., (2020, p. 1235), and Alsentzer et al., (2019, p. 74) showed that these models had a transformational influence on clinical decision-making, medical research, and healthcare management.

We preprocessed and thoroughly analyzed the medical transcription dataset to extract useful insights and build efficient NLP models. In the long run, these models may improve patient care and clinical outcomes by driving innovation in healthcare informatics.

4.3.2 Model Performance Analysis

When comparing traditional ML models with transformer-based models like BERT, it becomes evident that there are significant disparities in performance. The distinct structures and capabilities of these models may account for these discrepancies. Conventional machine learning models, such as Random Forest, SVM, XGBoost, and Logistic Regression, heavily depend on feature engineering and text representation techniques, such as TF-IDF. Although these models may be helpful in some situations, they face difficulties in capturing the intricate and complex contextual connections that are inherent in medical literature. Moreover, the following paragraphs enumerate the factors contributing to variances in performance.

a) Feature Engineering vs. Contextual Understanding

Traditional machine learning models are based on manual feature engineering, which can only identify surface-level patterns in text data. Techniques such as TF-IDF increase performance by considering the importance and rarity of words, but they fail to recognize the context in which words are employed. Medical language, for example, sometimes has special meanings that vary depending on context, which conventional models cannot fully understand. By contrast, BERT models with a deep bidirectional design that specializes in collecting such contextual details have the ability to examine both past and present terms in a phrase, which enables a more complete understanding and improves performance in tasks requiring exact knowledge of medical terminology.

b) Handling of High-Dimensional Data

High-dimensional data presents another obstacle for standard ML methods. Although models like SVM and Random Forest can somewhat manage high dimensions, they are challenging to understand and usually need large processing resources. Developed for large-scale, high-dimensional text data, BERT models rapidly process and extract significant patterns from massive medical corpora. To effectively manage long-range connections within text, their transformer design leverages self-attention approaches. This is particularly important for medical transcription tasks where the context spans many phrases or paragraphs.

c) Domain-Specific Pre-Training

Large volumes of domain-specific data taken from medical records and scientific publications have already been used in training the specialized forms of BERT, BioBERT, ClinicalBERT, and RoBERTa. Thanks to this pre-training, these models have great awareness of medical language and context-specific vocabulary. One such excellent example is BioBERT, which performs better than general-purpose models largely because it is pre-trained on PubMed abstracts, therefore addressing the linguistic patterns seen in biomedical literature. Standard ML models lose this domain-specific advantage as they cannot profit from significant pre-training.

d) Adaptability and Transfer Learning

The medical text categorization task finds BERT models especially helpful as they are rather flexible and may be modified for particular uses. Their ability to translate acquired knowledge from vast pre-trained datasets to specific applications reduces the requirement for sometimes uncommon in the medical sector massive, tagged datasets. The flexibility means that BERT models perform well even with insufficient task-specific training data, addressing one of the key constraints of standard ML techniques.

The observed performance trends correlate with previous studies on BERT and its different subtypes in the medical field. The studies show that BERT-based models yield higher performance than the traditional ML approaches due to better contextual understanding and transfer learning. In particular,

Lee et al. (2020, p. 1238) showed that BioBERT was much better than the baseline models when they tested it on biomedical named entity identification and relation extraction tasks. This study supports the findings that implementing domain-specific pre-training on biomedical texts provides learners with a significant advantage when navigating the complexities of medical terminology. Huang et al. (2020, p. 4) further revealed that ClinicalBERT, a BERT fine-tuned on clinical notes, outperformed normal BERT models on tasks related to hospital readmission and mortality rates, thereby reinforcing the domain adaptation work. In other words, the present study confirms prior research findings based on BERT models' consistently superior performance across multiple medical NLP applications, which speaks to the models' ability to decipher contextually rich medical texts.

BERT and its variants have greatly enhanced the generalizability and accuracy of medical transcription categorization. By addressing critical issues with clinical decision-making, this breakthrough showcases the game-changing potential of sophisticated deep learning models in healthcare. There will be better patient care methods and more accurate diagnostic tools through deep learning architectures' ability to manage complex medical data and extract significant insights, as demonstrated by the performance disparities between BERT and standard ML models.

4.3.3 Explainable AI Insights and Trustworthiness

a) SHAP Interpretation: The results of the SHAP analysis enhance the transparency and reliability of the BioBERT model in the classification of medical text transcriptions. SHAP contributes to the understanding of the decision-making process of complicated AI models by offering simple, visual explanations of which features (words or phrases) influence the model's predictions. In the medical industry, being transparent is especially important, as clinical validation and trust building among healthcare professionals depend on an awareness of the reasoning behind a categorization choice.

- Transparency: The SHAP values provide a precise breakdown of how each word or phrase contributes to the final prediction. This level of detail allows doctors to understand why the model reached a particular conclusion, improving its transparency.
- Trustworthiness: SHAP values promote trust with medical professionals by revealing which elements of the text have the most influence on decision-making processes. Physicians are more likely to trust and rely on AI-powered insights in their decision-making processes when they can see and grasp the logic behind the projections of a model.
- Accountability: Visualizing and quantifying the influence of certain words helps one to understand and take responsibility. Assessing the reliability and accuracy of AI models in important uses, such as healthcare, depends on stakeholders knowing how certain inputs generate specific outcomes.
- Error Analysis and Debugging: SHAP helps find any model errors or biases. For instance, if the model heavily relies on meaningless words, it may indicate a need for improvement in either the

model or the training data. This capacity is critical in improving model quality and increasing predictability.

- Improved Model Understanding: SHAP values provide patterns and insights that may not be evident from raw data or model outputs alone. This deeper understanding may serve as a guide for future model development and data gathering strategies, resulting in ongoing improvements in model performance.

b) Clinical Adoption and Application:

In practical clinical decision-making, the insights provided by SHAP can be transformative:

- Enhanced Clinical Validation: Clinicians can validate AI predictions by analyzing each word or sentence's contributions. This transparency ensures that the AI system is consistent with clinical thinking and norms.
- Trust Building with Clinicians: SHAP values assist clinicians in building trust by providing clear, visual explanations, which makes them more willing to embrace and rely on AI systems that they understand and trust.
- Improved Decision-Making: SHAP helps physicians identify the crucial parts of the medical text model, enabling them to make better decisions. Knowing why a particular model predicts a given diagnosis may help doctors validate and even modify their designated courses of action based on the model's calculations.
- Identifying and Mitigating Biases: When SHAP uncovers the overemphasis of certain minor features in prediction-making, it reveals model bias. This understanding allows model adjustments to decrease such biases, providing a better and fairer option.
- Educational Tool: Healthcare practitioners can learn how to use AI in healthcare through the SHAP visualizations, which show how AI models arrive at conclusions and the factors they consider important.

The results obtained from this study correspond with previous studies on the application of XAI and advanced language models in the healthcare discipline and contribute to the development of this scientific field. Thus, various empirical research studies have identified model interpretability and transparency.

In critical areas like healthcare, models that perform well while still being highly interpretable are essential Guo et al., (2016, p. 1). Supporting this assertion, our research shows that SHAP values greatly enhance the interpretability of models based on BioBERT, which in turn makes decision-making processes more transparent. In a similar way, Ong et al. (2010, p. 1) assessed the development of AI-powered clinical event report automated classifiers, emphasizing the need for transparency in such decisions. To satisfy the demand for transparency put forth by Ong et al. (2010, p. 1), our study uses SHAP to provide clear and detailed insights into model predictions. Furthermore, Shao et al. (2018, p. 2876) evaluated several text classification criteria, highlighting the significance

of reliability in therapeutic contexts. Our findings support this viewpoint, demonstrating that SHAP values can increase clinician trust by explaining the AI's decision-making process in a straightforward and intelligible manner. Finally, Chai et al. (2013, p. 982) demonstrated the relevance of recognizing and managing biases in AI systems using statistical text categorization of health information technology mishaps. Our study builds upon previous research by utilizing SHAP to detect and address potential biases in medical text classification. This approach enhances the transparency and dependability of AI systems in the healthcare field.

Integrating SHAP with BioBERT improves the transparency, trustworthiness, and overall interpretability of medical text transcription classification models. This integration not only improves model comprehension and stakeholder communication, but it also helps with error analysis and debugging, ensuring that AI models are correct, interpretable, and trustworthy for use in real-world healthcare settings.

4.4 Summary

The study using medical transcription data demonstrates the revolutionary influence of enhanced pre-processing, complex ML models, and XAI techniques on data classification and comprehension. Rigid preprocessing, including normalization, noise reduction, and keyword analysis, reduced the dataset to a balanced and therapeutically appropriate format. Traditional ML models performed rather well; however, transformer-based models, like BioBERT, outperformed them dramatically. This emphasizes the need for domain-specific architectures for efficiently digesting specialized medical texts. The use of SHAP values to determine model explainability provides unambiguous insights into feature importance, which is critical for clinical validation and acceptance. This transparency guarantees that AI-driven insights are not only accurate but also interpretable and trustworthy, improving communication with non-technical stakeholders and enabling practical clinical decision-making. These enhancements demonstrate the potency of merging sophisticated NLP and AI, resulting in models that are resilient, dependable, and beneficial in clinical environments. This can lead to better healthcare outcomes and big steps forward in medical informatics.

Chapter 5: Conclusion

In this work, we thoroughly discussed the efficiency and interpretability of BERT-based language models for medical text transcription categorization. The study aimed to address the issue of complex NLP models' black box nature by applying XAI techniques, particularly SHAP. The objective was to enhance the use of these models in clinical settings. The main goal was to assess the progress of these models to promote transparent decision-making and improve accuracy. This would ultimately enhance the reliability and effectiveness of AI applications in the healthcare sector. The study also explored how SHAP could provide valuable insights into model predictions, which could enhance the trust of healthcare professionals. The research also looked at what would happen if these models were used in real-life clinical settings. It showed that these models could help connect the gap between advanced AI capabilities and real-world use in the healthcare sector.

5.1 Summary of Findings

This section provides a concise overview of the main discoveries made in the preceding chapters of the thesis. These chapters focused on assessing and incorporating BERT-based language models with XAI approaches to improve the precision and transparency of medical text transcription.

Chapter 1 highlighted the importance of accurate medical text transcription for effective patient care and well-informed clinical decisions. The study analyzed the improvements in medical text classification achieved by advanced NLP models such as BERT while simultaneously highlighting their lack of transparency. Transparency is critical in therapeutic settings because it is required to comprehend model decisions. The chapter presented XAI approaches, namely SHAP, as a method to improve the interpretability of models and cultivate trust among healthcare practitioners. The chapter provided an overview of the study goals and objectives, which include assessing BERT-based models and using XAI to tackle the inherent black box nature of these models.

Chapter 2 included a thorough summary of the current research on transcription of medical text data and the integration of BERT with XAI methods. The research work discussed the progression of medical text transcription, starting with conventional techniques and progressing to more sophisticated NLP models. The study emphasized the importance of maintaining consistency and precision in classification. The chapter investigated the advances made by deep learning models such as BERT, BioBERT, ClinicalBERT, and RoBERTa and compared the performance of conventional machine learning methods. The investigation also covered the shortcomings of conventional approaches, underlined the need for transparency in AI models, and explored the possibilities of XAI techniques to solve these difficulties. The evaluation highlighted BERT's advantages in understanding medical language and the need for XAI to improve the dependability and transparency of findings.

Chapter 3 outlined the approach to analyze BERT-based models and XAI methods. The study strategy was outlined, including the steps of objective definition, data collection via mtsamples.com, and

preprocessing. This chapter detailed the steps used to compare several models, including more conventional ML algorithms and BERT-based ones such as BioBERT, ClinicalBERT, and RoBERTa. One important aspect of XAI was the introduction of SHAP, which aimed to improve model transparency. The method also included testing for robustness and validity, as well as addressing limitations and ethical concerns such as data bias and privacy concerns. A blend of state-of-the-art XAI methods, thorough data analysis, and model validation led to improved accuracy and transparency.

Chapter 4 discussed the study's findings, which included evaluating models and XAI approaches. It provides a concise overview of the dataset's characteristics, the preprocessing methods used, and the comparative performance of conventional ML techniques in contrast to BERT-based models. We identified BioBERT as the most efficient model, outperforming standard approaches by a wide margin in classification tasks. The chapter emphasized the efficacy of SHAP in offering discernment into model choices, augmenting transparency, and bolstering trust. The chapter emphasized SHAP's ability to clarify predictions, identify errors, and improve the model. The findings proved that the contextual and domain-specific pre-training of BERT-based models, such as BioBERT, allows them to understand complex medical texts. By including SHAP, we were able to address issues with the black box nature of advanced NLP models and make the model more interpretable.

This thesis demonstrates that while sophisticated NLP models like BioBERT exhibit exceptional performance in medical text categorization, the incorporation of XAI approaches like SHAP is essential for improving model transparency and confidence. The study emphasizes the significance of achieving a balance between accuracy and interpretability to enhance the practical implementation of healthcare solutions. This, in turn, will contribute to more dependable and efficient patient care, as well as clinical decision-making.

5.2 Potential Implications

The use of sophisticated NLP models such as BERT, BioBERT, ClinicalBERT, and RoBERTa, together with XAI techniques like SHAP, in the categorization of medical text transcription has noteworthy consequences for the healthcare sector. These findings have implications for several domains, including clinical practice, healthcare administration, patient care, and the wider realm of AI in medicine.

a) Enhanced Clinical Decision-Making: Integration of advanced NLP models such as BERT, BioBERT, ClinicalBERT and RoBERTa with medical text transcription greatly enhances clinical decision-making. These models' tremendous accuracy in recognizing medical language ensures that healthcare professionals get precise and correct information, reducing the possibility of errors. Furthermore, improving the quality of the knowledge gained from medical text data is the rich contextual awareness of medical language and nuances given by models such as BioBERT (Lee et al., 2020, p. 2). This technique generates more informed and precise clinical decisions, which helps to improve the general standard of treatment quality and patient outcomes.

b) Increased Trust and Adoption of AI in Healthcare: XAI techniques as SHAP help AI models to be much more transparent and interpretable. This solves the issue of conventional AI models being black box and provides a clear justification for the prediction process, which is critical for prediction confidence among medical practitioners. Healthcare organizations will widely use and accept AI if they can understand and justify its outcomes. Furthermore, transparency promotes accountability by empowering medical practitioners to make informed AI decisions, thereby facilitating the ethical and efficient application of AI in healthcare.
c) Error Analysis and Continuous Improvement: The use of SHAP in interpreting models' predictions helps identify and rectify the sources of biases and individual errors. This capability is critical for improving and optimizing AI models. Thus, recognizing and solving these problems will allow us to optimize AI systems in healthcare facilities and provide patients with more equal and accurate treatment. Moreover, SHAP enables better debugging and optimization processes through its functions of visualization and interpretation of model decisions, thus promoting the creation of higher-quality AI. This constant loop of improvement is critical when it comes to the healthcare applications of AI (Lundberg & Lee, 2017, p. 5).
d) Patient Care and Safety: Medical text transcription inevitably impacts patients, particularly when care and safety are at risk. Continuity of care and continuation of patient treatment rely heavily on comprehensive patient documentation, which, in turn, is only possible with higher transcription competency. Correct categorization of medical text data improves patient outcomes and safety by reducing the likelihood of clinical errors. Healthcare practitioners can make better decisions with the consistent and trustworthy information offered by modern NLP models. This, in turn, leads to better patient care and safety.
e) Efficiency and Productivity in Healthcare Management: The use of advanced NLP models to enhance medical documentation improves automation and productivity in healthcare management. Consequently, these models offer valuable patient information while reducing the workload for care providers in digital documentation. Further, the greater accuracy in the medical text classification makes it possible for administrations in different healthcare facilities to allocate resources to the right priority areas in their organizations. This efficiency in resource use, therefore, automatically enhances the efficiency and effectiveness of the provision of health care services (Mujtaba et al., 2019, p . 512).
f) Broader Implications for AI in Medicine: The integration of XAI with BERT-based models has been a successful attempt, opening new avenues in medical AI applications. It demonstrates the feasibility and utility of combining explainability with NLP, a technologically advanced feature, thus promoting additional research and progression in this area. This research emphasizes the importance of transparency, fairness, and accountability in developing and implementing AI technologies so that they do not harm health care services through improper use. By addressing these significant aspects, the study lays the groundwork for considerably more reliable and efficient AI implementations in clinical practice and contributes to the advancement of AI in medicine.

Thus, the integration of BERT-based models with XAI techniques in medical text transcription can have a tremendous impact. Besides, this research contributes not only to the development of state-of-the-art AI for healthcare but also to building more transparent and trustworthy AI for clinical use. By assessing the methods of accuracy and interpretability and by presenting the results of this study, this research lays down the foundation for enhancing patient healthcare and paving the way towards perfect potent medicine with the help of innovative technological tools.

5.3 Limitations of the Study

While this research provides valuable insights into the performance and transparency of BERT-based models for medical text transcription classification, several limitations need to be acknowledged:

a) Dataset Limitations:

- Sample Size: This research uses an extensive dataset, totaling 4,999 records. After preprocessing and using domain knowledge, we further reduce this dataset to 2,324 records. Even though this is a sufficient sample for a preliminary study, it may not capture the whole range of medical transcription services. A larger dataset would yield a more comprehensive and robust assessment.
- Data Quality: Missing values and inconsistencies are some of the most critical problems of datasets that require careful preparation. However, there are basic quality problems that may influence the models' performance and applicability even after cleaning and normalizing the data.

b) Model Generalizability:

- Domain-Specific Training: Although models like BioBERT and ClinicalBERT have demonstrated better performance because of the domain-specific pre-training, it might be the case that these models are effective only for the types of medical text they have been trained on. They might perform differently when used in other forms of medical documents or in other settings in clinical practice.
- Overfitting: Even applying cross-validation and the SMOTE method, there will always be a possibility that there is overfitting, more so with a complicated model such as BERT. They include overfitting, which can reduce the model’s ability to generalize from the training data to other unseen data.

c) Explainability Challenges:

- Complexity of SHAP Values: While using SHAP makes the models more interpretable, understanding SHAP values for a complicated model such as BERT is not an effortless task. The interpretation of model decisions can be challenging due to the high dimensionality and the nature of interactions within the data that are difficult to comprehend by non-technical personnel.

- Scalability of XAI Techniques: The application of XAI techniques such as SHAP to sophisticated models and large datasets is computationally intensive. This may restrict their practical implementation in real-time clinical settings, where prompt decision-making is essential.

d) Ethical and Practical Considerations:

- Data Privacy: In the medical field, the confidentiality and security of information are one of the most critical and sensitive matters. Even though this study anonymized the data, privacy laws and ethical concerns restrict the practical applications of the developed models.
- Bias and Fairness: Despite some attempts to mitigate the biases, the training datasets may still bias the models. This could result in biased or unfair decisions, which pose a significant risk, particularly in a health-related system where the outcome directly impacts the patient.

e) Technical Constraints:

- Computational Resources: Fine-tuning and training large models such as BERT also demand a lot of computational power, especially from GPUs. This can prove to be a drawback for small-scale institutions or research groups who might not be able to afford to acquire the necessary facilities.
- Implementation Challenges: Despite the recent developments in the field of NLP and the integration of explainability, the implementation of these models and approaches into clinical practice raises numerous technical problems. This also involves EHR integration and other aspects of the AI system's dependability and robustness in the context of rapidly evolving clinical practices.

f) Validation and Clinical Utility:

- Limited Clinical Validation: Although the study has positive outcomes, the evaluation of the models and techniques necessitates further clinical assessment for their applicability and efficiency in clinical practice. Some pilot studies and clinical trials are required to validate and prove the real-world applicability and effectiveness of these AI systems in healthcare.

It's important to note that this study adds to the existing body of literature by explaining BERT-based models and using XAI techniques. However, these problems need to be fixed to encourage the practical use and adoption of these technologies in the clinical setting. To increase the robustness and application value of AI-driven medical text transcription systems, further work should focus on collecting larger volumes of experiment data, improving methods for explaining results, and conducting thorough clinical validations to enhance the robustness and utility of AI-driven medical text transcription systems.

5.4 Recommendations and Future Research

Based on the findings and limitations of this study, several recommendations and areas for future research are proposed to further enhance the performance and interpretability of BERT-based models in medical text transcription classification:

a) Integrating Various XAI Techniques:

- Attention Mechanisms: In the case of employing attention mechanisms in the BERT model, such mechanisms improve interpretations as they create a way to show the model's details in the text as it computes results. This computation could help explain why the model arrived at a certain decision; this makes comprehension easier.
- LIME: It is possible that the integration of LIME with SHAP will give a better explanation of the model's predictions. LIME's method, which involves producing simpler and more understandable models on a local level for approximation, may complement SHAP's global results (Lundberg & Lee, 2017, p. 9). All of this, in combination, allows for a deeper understanding of the model's behavior

b) Comparative Evaluation with Human Experts:

- Benchmarking Against Human Classification: Setting up a baseline to measure the model's efficiency and standards involves comparing the classifications generated by the BERT-based model with those generated by trained medical practitioners (Talebi et al., 2024, p. 7). Such a comparison can reveal how the model performs relative to the human experts and define where the model agrees or disagrees with the human experts
- Detailed Error Analysis: An extensive error analysis can be performed by comparing the results of the models with the classifications provided by the human practitioner to identify systematic errors. This can help inform better improvements on the model and its training process, making it better and having fewer errors.

c) Validation of Explanations Through Stress Tests:

- Erasure Method: We can design stress tests using the erasure method to evaluate the XAI techniques' explanations. The process involves gradually eliminating portions of the input text that the model considers significant and pertinent and observing the impact on the model's result. (Talebi et al., 2024, p. 5). It will also show whether the identified features do indeed play a role in the decision-making process of a given subject

d) Expanding and Diversifying Datasets:

- Larger and More Diverse Datasets: Future studies should aim at collecting an even larger and more diverse sample of medical text data, with the idea of enlarging the sample space and increasing the ability of the model to generalize. Different sets of data from several fields of medicine and subjects of different ages and genders will thus ensure that the models can predict and accept all sorts of medical texts and clinical situations.

e) Enhancing Model Generalizability:

- Domain Adaptation: To improve their performance, we can incorporate domain adaptation methods to fine-tune BERT models on various subdomains in the medical field. Future work should further explore transfer learning and domain-specific pre-training techniques to enhance the model's generality for distinct forms of medical documents.

f) Simplifying Explainability for Non-Technical Users:

- User-Friendly Explanations: While AI researchers primarily understand SHAP and other XAI outputs, analogous advances suggest that it is crucial to transform these techniques into packages that non-technical healthcare specialists can intuitively use. Introducing simple and clear interfaces and visualizations may help clinicians better understand the AI's predictions, thus increasing its usage.

g) Addressing Ethical and Practical Deployment Issues:

- Data Privacy and Security: It is critical to pay particular attention to the aspects of data protection and data security. Future research should focus on improving various methods for anonymizing medical data and maintaining proper data management procedures, particularly in scenarios related to realistic clinical environments.
- Bias and Fairness: Further work is still required to eliminate the possibility of biases in the training data for the models and in the predictions they provide. To predict and avoid the negative consequences of discriminatory effects with respect to patient groups, approaches to introducing fairness and equality into the system of AI's application in healthcare are necessary.

h) Comprehensive Clinical Validation:

- Clinical Trials and Pilot Studies: Further research and pilot clinical tests and clinical trials will be required to prove the utility and accuracy of the BERT-based models in clinical practice. This will aid in establishing practical realism as well as safety for the aforementioned AI systems in health facilities.

i) Integration of Additional XAI Techniques:

- Combining SHAP with Other Techniques: To get a full picture of the model's prediction, more research should investigate how SHAP could be combined with other types of XAI, such as sensitivity analysis or feature importance analysis (H. Wang et al., 2024, p. 2). This combined strategy can provide more detailed knowledge about model behavior and decision-making processes.

If we adhere to these recommendations and pursue the suggested paths for future research, we can significantly advance the field of medical text transcription classification. We can also deduce that these endeavors will enhance the accuracy, exposition, and interpretability of BERT-based models, rendering them more applicable in clinical settings and aligning with the objectives of AI in healthcare projects.

Appendix A: Medical Text Data Visualizations and BERT Performance Output

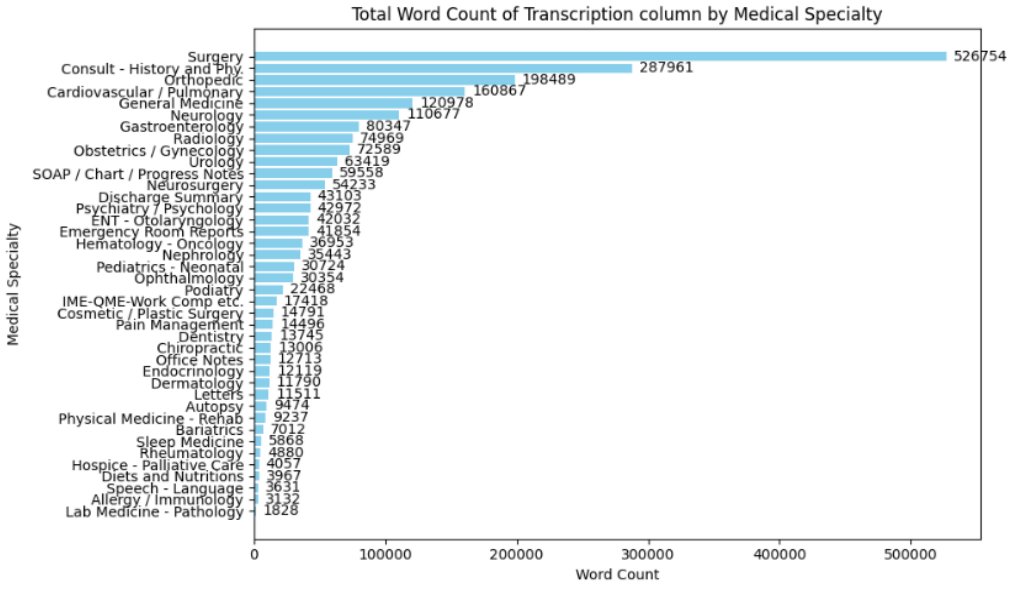
This section explains the visualization of additional information in the research methodology chapter, more precisely, in section 3.5, which indicates EDA. Word clouds, a simple and easy-to-implement visualization form, serve as a first overview. They usually present some of the frequently used terms from the text as a list of words with weights in a particular spatial orientation. The sizes of the words correspond respectively with their importance or occurrence, and the rest of the graphic features (such as color, location, and orientation) are corrected either for aesthetic reasons or in order to encode other information (Lohmann et al., 2015, p. 1). Figure 40 shows a word cloud graphic representation of the medical specialization column.

Figure 40. Word cloud of the medical specialty column

Source: Own results.

The output shown in Figure 41 is a horizontal bar chart that displays the total word count for the transcription column for each medical specialty, arranged in descending order. Additionally, Figure 42 presents the word cloud representation of the transcription data.

Figure 41. Total word count of transcription column by medical specialty



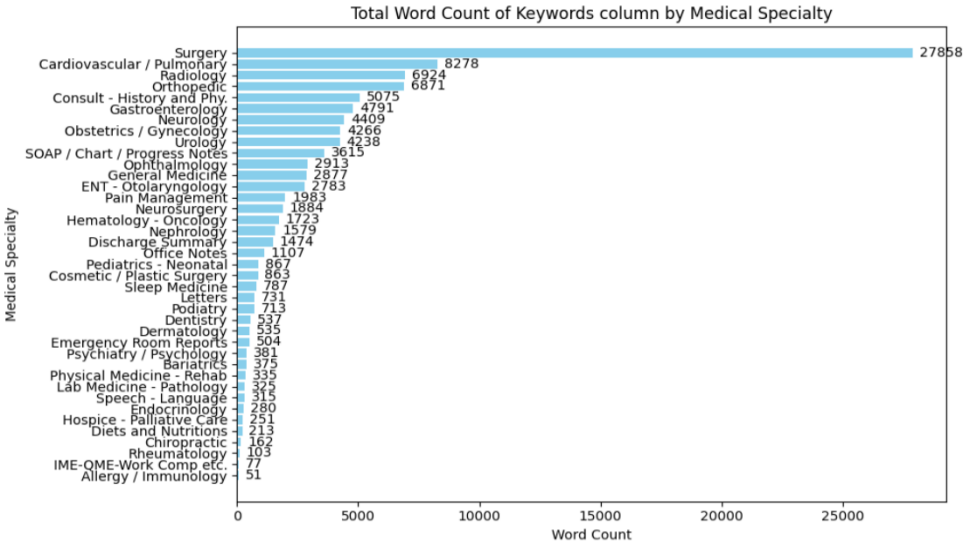
Source: Own results.

Figure 42. Word cloud of the transcription column

Source: Own results.

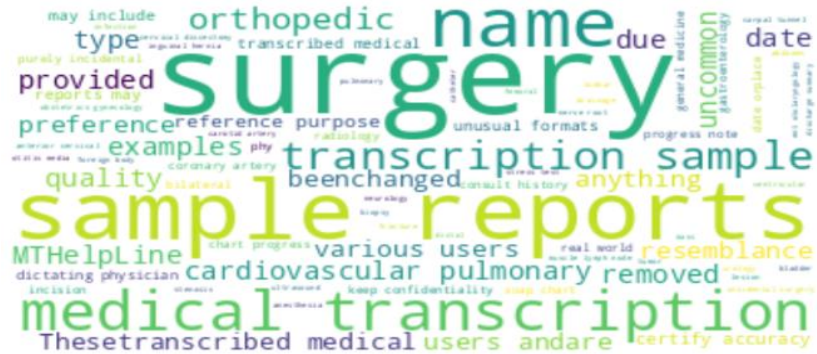
In addition, Figure 43 shows the bar chart on keyword word count by medical discipline, and Figure 44 shows the word cloud for the keyword column.

Figure 43. Total word count of keywords column by medical specialty



Source: Own results.

Figure 44. Word cloud of the keyword's column

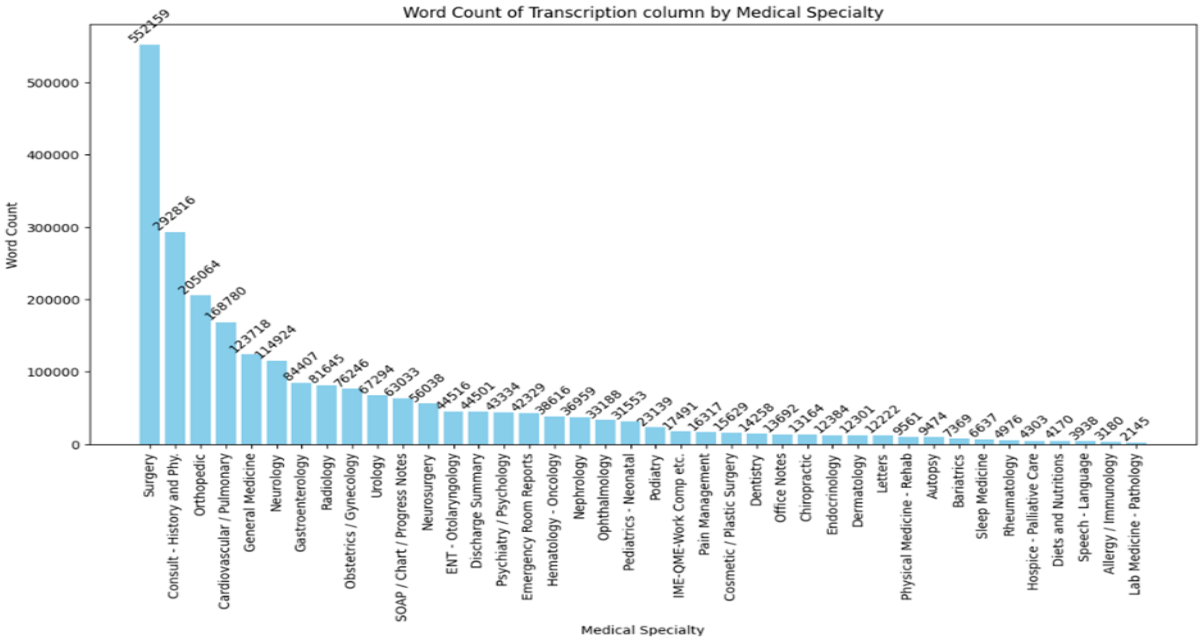


Source: Own results.

In the data preprocessing step, after concatenating the 'keywords' and 'transcription' columns and subsetting to include the appropriate columns ('transcription' and 'medical_specialty'), as described

in section 3.6.2, the word count by medical specialty is analyzed in section 3.6.3. The bar graph visualization shown in Figure 45 provides a clear depiction of the word count distribution, enabling straightforward comparison among medical specialties.

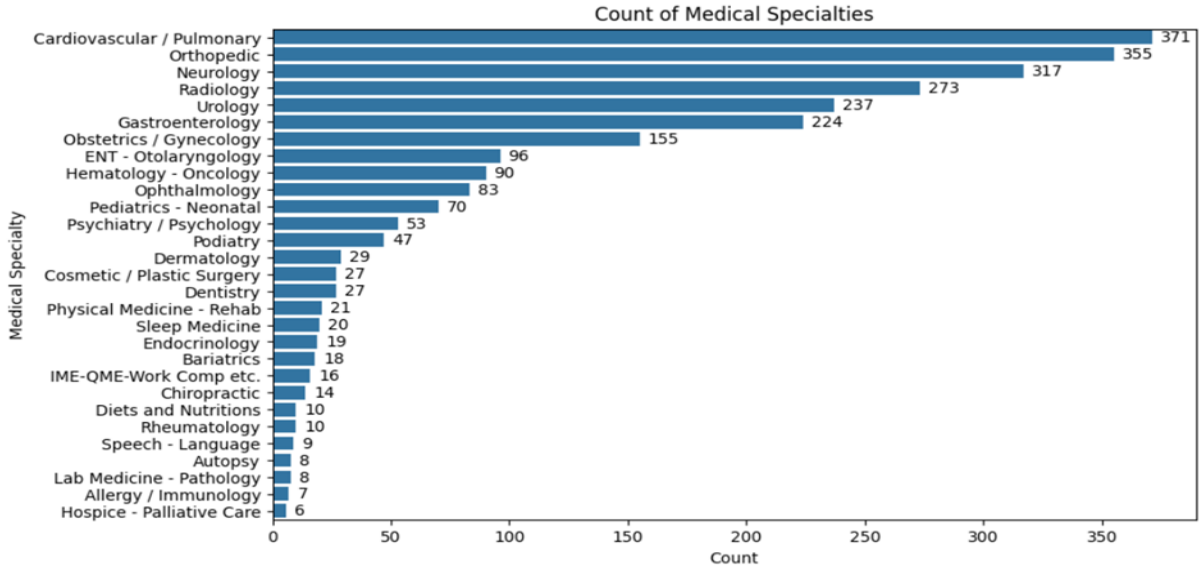
Figure 45. Updated word count of the transcription column by medical specialty after concatenation



Source: Own results.

Furthermore, after applying domain knowledge in section 3.6.6, Figure 46 presents a bar graph of the count of medical specialties that aids in visualizing this distribution more conveniently. This Figure 46 provides insight into the relative sizes of samples in several medical fields, pointing the way for future research and model development.

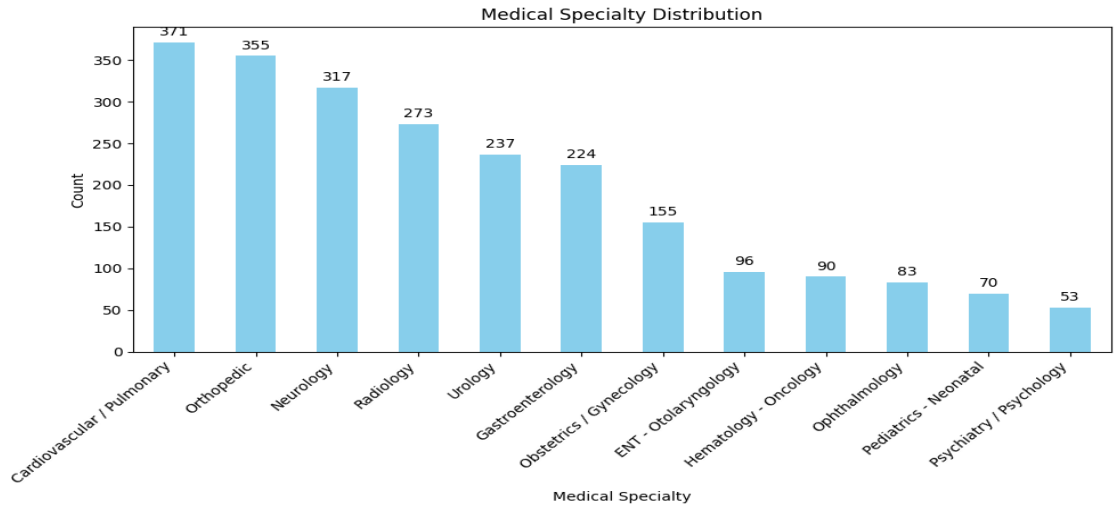
Figure 46. Count of medical specialties after applying domain knowledge



Source: Own results.

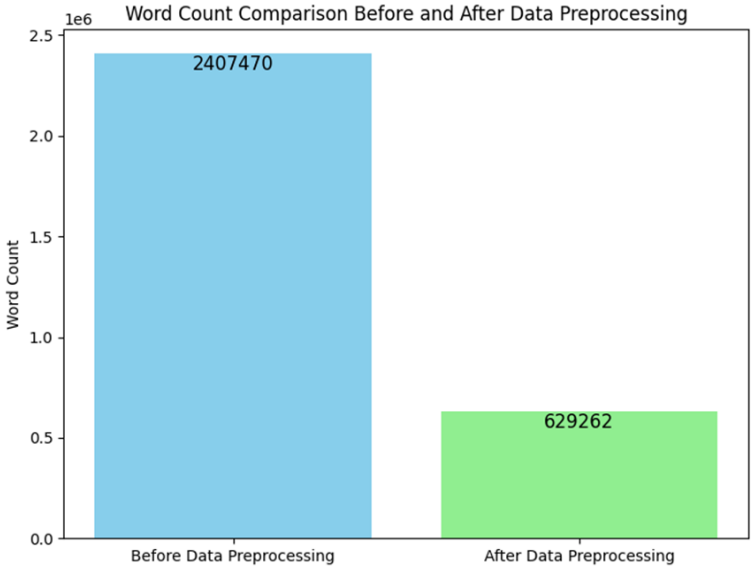
Figure 47 shows the distribution of medical specialties after incorporating domain knowledge and adjusting the categories. In section 3.6.7, categories with fewer than 50 samples were excluded to focus on those with a larger representation. This adjustment resulted in a reduction of the total word count by 73.86%. Figure 48 depicts a bar graph comparing the word count before and after the data preprocessing of the transcription column.

Figure 47. Bar graph of finalized medical specialty distribution



Source: Own results.

Figure 48. Word count comparison before and after data preprocessing of the transcription column data



Source: Own results.

Additionally, in section 3.8.2 of advanced model development, the performance of various BERT-based medical text categorization models (BioBERT, ClinicalBERT, and RoBERTa) is shown in Output 1 below.

Output 1. The performance of different BERT (BioBERT, ClinicalBERT and RoBERTa) medical text classification algorithm models

Classification Report for BERT:

	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.93	0.95	0.94	74
ENT - Otolaryngology	1.00	0.95	0.97	19
Gastroenterology	0.83	0.89	0.86	45
Hematology - Oncology	1.00	0.78	0.88	18
Neurology	0.81	0.89	0.85	63
Obstetrics / Gynecology	0.83	0.97	0.90	31
Ophthalmology	1.00	0.94	0.97	17
Orthopedic	0.91	0.94	0.92	71
Pediatrics - Neonatal	0.85	0.79	0.81	14
Psychiatry / Psychology	0.89	0.73	0.80	11
Radiology	1.00	0.85	0.92	55
Urology	0.93	0.91	0.92	47
accuracy			0.90	465
macro avg	0.92	0.88	0.90	465
weighted avg	0.91	0.90	0.90	465

Classification Report for BioBERT:

	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.95	0.93	0.94	74
ENT - Otolaryngology	1.00	1.00	1.00	19
Gastroenterology	0.92	1.00	0.96	45
Hematology - Oncology	1.00	0.83	0.91	18
Neurology	0.84	0.92	0.88	63
Obstetrics / Gynecology	0.86	0.97	0.91	31
Ophthalmology	1.00	1.00	1.00	17
Orthopedic	0.94	0.96	0.95	71
Pediatrics - Neonatal	0.92	0.79	0.85	14
Psychiatry / Psychology	0.89	0.73	0.80	11
Radiology	0.98	0.85	0.91	55
Urology	0.98	0.98	0.98	47
accuracy			0.93	465
macro avg	0.94	0.91	0.92	465
weighted avg	0.93	0.93	0.93	465

Classification Report for ClinicalBERT:

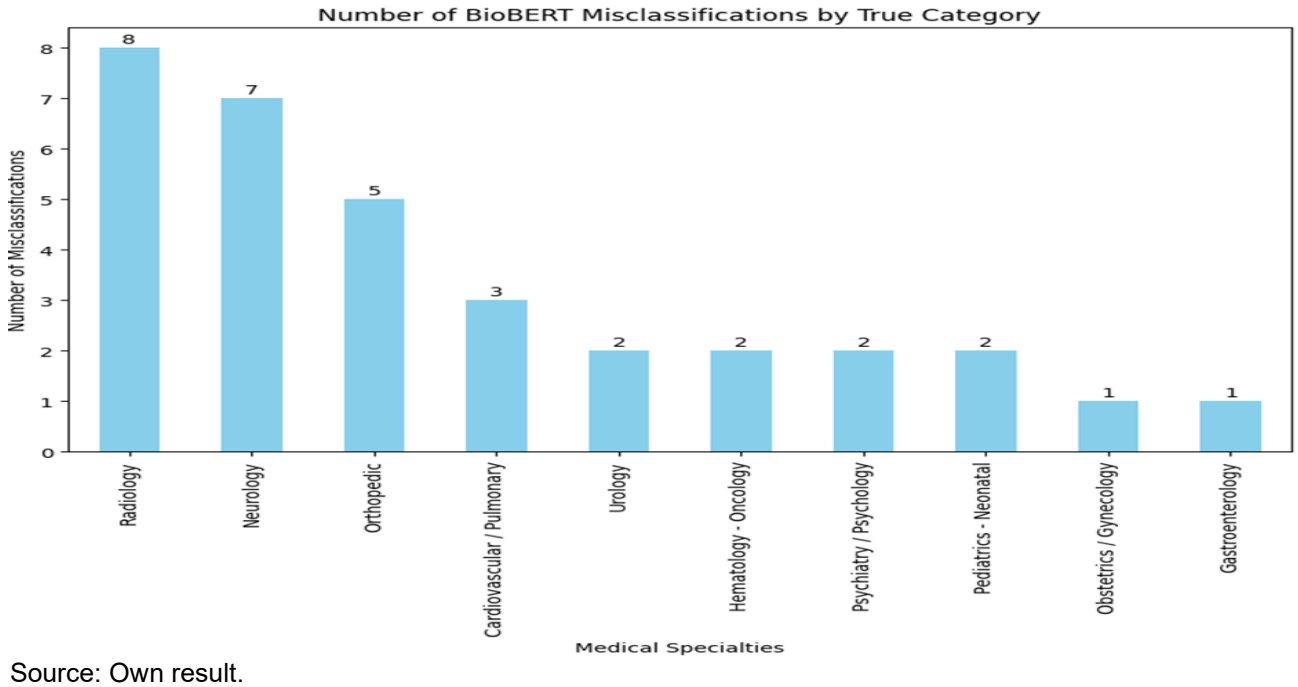
	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.93	0.92	0.93	74
ENT - Otolaryngology	0.95	1.00	0.97	19
Gastroenterology	0.96	1.00	0.98	45
Hematology - Oncology	0.89	0.89	0.89	18
Neurology	0.86	0.90	0.88	63
Obstetrics / Gynecology	0.88	0.97	0.92	31
Ophthalmology	1.00	0.94	0.97	17
Orthopedic	0.89	0.96	0.93	71
Pediatrics - Neonatal	0.77	0.71	0.74	14
Psychiatry / Psychology	0.89	0.73	0.80	11
Radiology	0.96	0.85	0.90	55
Urology	1.00	0.94	0.97	47
accuracy			0.92	465
macro avg	0.92	0.90	0.91	465
weighted avg	0.92	0.92	0.92	465

Classification Report for RoBERTa:

	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.93	0.95	0.94	74
ENT - Otolaryngology	0.95	1.00	0.97	19
Gastroenterology	0.98	0.98	0.98	45
Hematology - Oncology	0.88	0.83	0.86	18
Neurology	0.83	0.87	0.85	63
Obstetrics / Gynecology	0.86	0.97	0.91	31
Ophthalmology	1.00	0.94	0.97	17
Orthopedic	0.91	0.94	0.92	71
Pediatrics - Neonatal	0.92	0.86	0.89	14
Psychiatry / Psychology	0.90	0.82	0.86	11
Radiology	0.96	0.85	0.90	55
Urology	0.96	0.91	0.93	47
accuracy			0.92	465
macro avg	0.92	0.91	0.92	465
weighted avg	0.92	0.92	0.92	465

Furthermore, the misclassification error analysis of the BioBERT model, which outperforms the other BERT models, is illustrated in Figure 49. This figure depicts the number of BioBERT misclassifications by true category as a bar graph.

Figure 49. Number of BioBERT misclassifications by true category

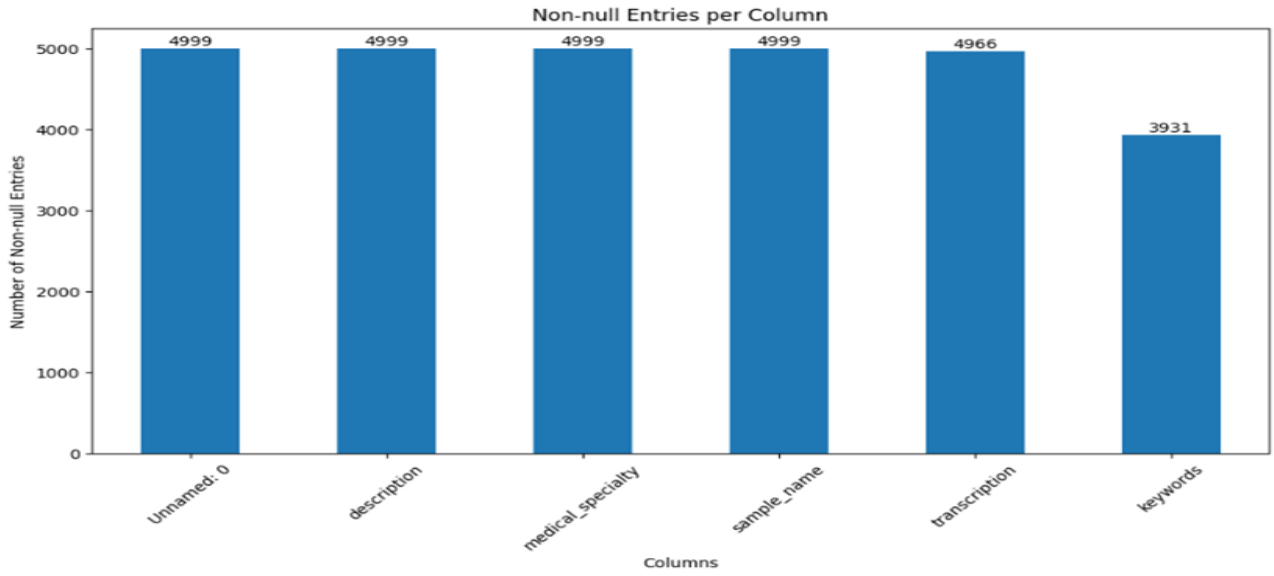


Source: Own result.

Appendix B: Dataset Characteristics and Preprocessing

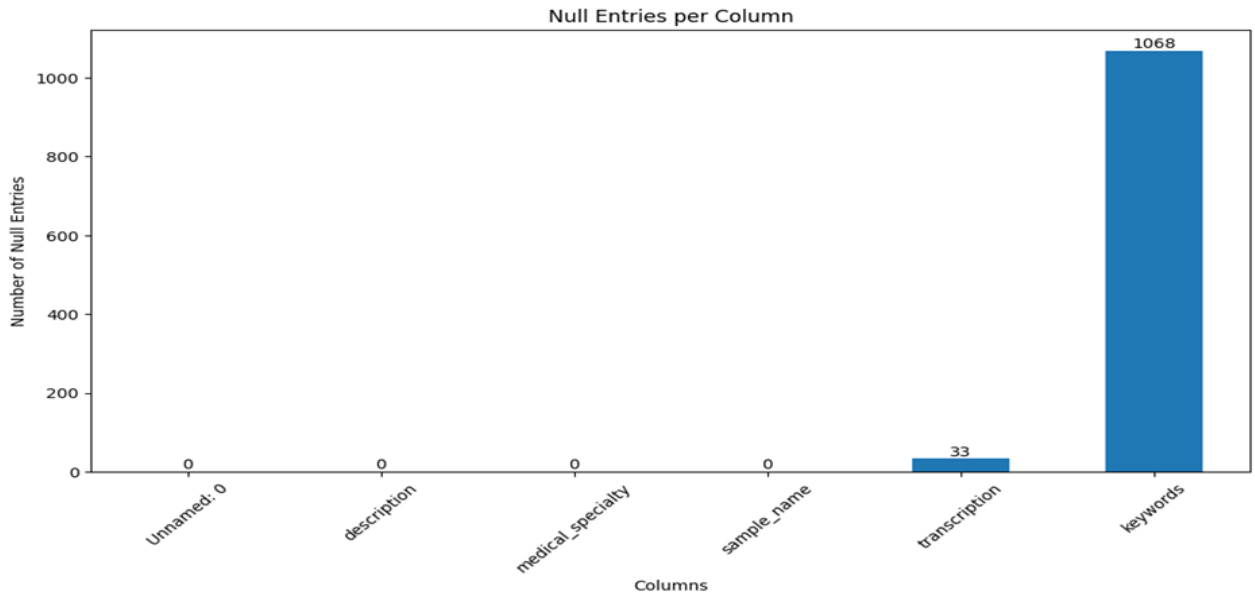
In this section, additional information from the research findings chapter is illustrated. In section 4.2.1, dataset characteristics and preprocessing, the bar graph in Figure 50 below shows the number of non-null value entries per column, while another Figure 51 displays the number of null value entries per column in the medical transcription dataset.

Figure 50. Bar graph of number of non-null values entries per each column



Source: Own result.

Figure 51. Bar graph of number of null values entries per each column



Source: Own result.

Furthermore, Table 15 below provides a clearer perspective on the prevalence of different specialties, facilitating targeted analysis and enhancing model accuracy. This is achieved by applying domain knowledge to refine and adapt the dataset's categories, which is particularly useful in medical text classification applications.

Table 15. Count of medical specialties after applying domain knowledge

Medical Specialty	Count
Cardiovascular / Pulmonary	371
Orthopedic	355
Neurology	317
Radiology	273
Urology	237
Gastroenterology	224
Obstetrics / Gynecology	155
ENT – Otolaryngology	96
Hematology – Oncology	90
Ophthalmology	83
Pediatrics – Neonatal	70
Psychiatry / Psychology	53
Podiatry	47
Dermatology	29
Cosmetic / Plastic Surgery	27
Dentistry	27
Physical Medicine – Rehab	21
Sleep Medicine	20
Endocrinology	19
Bariatrics	18
IME-QME-Work Comp etc.	16
Chiropractic	14
Rheumatology	10
Diets and Nutritions	10
Speech – Language	9
Lab Medicine – Pathology	8
Autopsy	8
Allergy / Immunology	7
Hospice - Palliative Care	6

Source: Own result.

Appendix C: GitHub Repository

The GitHub repository for the Python code implementation of the above research work can be accessed at the following link:

GitHub Repository: Master-Thesis-Transparency_MedTranscription_BERT_XAI

The repository consists of the following folders and files:

1. Medical Transcription Data (CSV file)

- 'mtsamples.csv'

2. Python code - ML model with XAI (PDF file)

- 'Medical Text Transcription Classification ML model with XAI Integration.pdf'

3. Python code - ML model with XAI (ipynb file)

- 'Medical Text Transcription Classification ML model with XAI Integration.ipynb'

4. README.md

This repository contains the dataset used in the study of research work, the machine learning model code, and an explanation of how Explainable AI approaches were integrated. The README.md file in the repository contains additional information on these files.

Bibliography

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity Checks for Saliency Maps. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html

Al Kuwaiti, A., Nazer, K., Al-Reedy, A., Al-Shehri, S., Al-Muhanna, A., Subbarayalu, A. V., Al Muhanna, D., & Al-Muhanna, F. A. (2023). A Review of the Role of Artificial Intelligence in Healthcare. *Journal of Personalized Medicine*, *13*(6), Article 6. https://doi.org/10.3390/jpm13060951

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera Triguero, F. (2023). *Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence*. https://doi.org/10.1016/j.inffus.2023.101805

Almazaydeh, L., Abuhelaleh, M., Tawil, A. A., & Elleithy, K. (2023). Clinical Text Classification with Word Representation Features and Machine Learning Algorithms. *International Journal of Online and Biomedical Engineering (iJOE)*, *19*(04), Article 04. https://doi.org/10.3991/ijoe.v19i04.36099

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., & McDermott, M. (2019). Publicly Available Clinical BERT Embeddings. In A. Rumshisky, K. Roberts, S. Bethard, & T. Naumann (Eds.), *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (pp. 72–78). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-1909

Asudani, D. S., Nagwani, N. K., & Singh, P. (2023). Impact of word embedding models on text analytics in deep learning environment: A review. *Artificial Intelligence Review*, *56*(9), 10345–10425. https://doi.org/10.1007/s10462-023-10419-1

Babu, A., & Boddu, S. B. (2024). BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding. *Exploratory Research in Clinical and Social Pharmacy*, *13*, 100419. https://doi.org/10.1016/j.rcsop.2024.100419

Bajwa, J., Munir, U., Nori, A., & Williams, B. (2021). Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthcare Journal*, *8*(2), e188–e194. https://doi.org/10.7861/fhj.2021-0095

Barr, P. J., Dannenberg, M. D., Ganoe, C. H., Haslett, W., Faill, R., Hassanpour, S., Das, A., Arend, R., Masel, M. C., Piper, S., Reicher, H., Ryan, J., & Elwyn, G. (2017). Sharing Annotated Audio Recordings of Clinic Visits With Patients—Development of the Open Recording

Automated Logging System (ORALS): Study Protocol. *JMIR Research Protocols*, *6*(7), e7735. https://doi.org/10.2196/resprot.7735

Barragán-Montero, A., Bibal, A., Dastarac, M. H., Draguet, C., Valdés, G., Nguyen, D., Willems, S., Vandewinckele, L., Holmström, M., Löfman, F., Souris, K., Sterpin, E., & Lee, J. A. (2022). Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency. *Physics in Medicine & Biology*, *67*(11), 11TR01. https://doi.org/10.1088/1361-6560/ac678a

Basystiuk, O., & Melnykova, N. (2022, December 16). *Multimodal Approaches for Natural Language Processing in Medical Data*.

Bhandari, S. (n.d.). *Medical Transcription: Everything You Need to Know and More*. Retrieved July 23, 2024, from https://reduct.video/blog/medical-transcription/

Bharati, S., Mondal, M. R. H., & Podder, P. (2024). A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When? *IEEE Transactions on Artificial Intelligence*, *5*(4), 1429–1442. IEEE Transactions on Artificial Intelligence. https://doi.org/10.1109/TAI.2023.3266418

Binder, M., Heinrich, B., Hopf, M., & Schiller, A. (2022). Global reconstruction of language models with linguistic rules – Explainable AI for online consumer reviews. *Electronic Markets*, *32*(4), 2123–2138. https://doi.org/10.1007/s12525-022-00612-5

Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. *Journal of Data Analysis and Information Processing*, *8*(4), Article 4. https://doi.org/10.4236/jdaip.2020.84020

Boyko, N., & Boksho, K. (2020). *Application of the Naive Bayesian Classifier in Work on Sentimental Analysis of Medical Data*. International Workshop on Informatics & Data-Driven Medicine. https://www.semanticscholar.org/paper/Application-of-the-Naive-Bayesian-Classifier-in-on-Boyko-Boksho/d57726a82e28ccee12c8d9c9d07f913ddc4f2dd0

Brnabic, A., & Hess, L. M. (2021). Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. *BMC Medical Informatics and Decision Making*, *21*(1), 54. https://doi.org/10.1186/s12911-021-01403-2

Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of Explainable AI Techniques in Healthcare. *Sensors (Basel, Switzerland)*, *23*(2), 634. https://doi.org/10.3390/s23020634

Chai, K. E. K., Anthony, S., Coiera, E., & Magrabi, F. (2013). Using statistical text classification to identify health information technology incidents. *Journal of the American Medical Informatics Association: JAMIA*, *20*(5), 980–985. https://doi.org/10.1136/amiajnl-2012-001409

Chakrobartty, S., & El-Gayar, O. (2021). *Explainable Artificial Intelligence in the Medical Domain: A Systematic Review*. Americas Conference on Information Systems. https://www.semanticscholar.org/paper/Explainable-Artificial-Intelligence-in-the-Medical-Chakrobartty-El-Gayar/4e405e87a55611328d2f408ae9d4164373454296

Chen, P.-F., He, T.-L., Lin, S.-C., Chu, Y.-C., Kuo, C.-T., Lai, F., Wang, S.-M., Zhu, W.-X., Chen, K.-C., Kuo, L.-C., Hung, F.-M., Lin, Y.-C., Tsai, I.-C., Chiu, C.-H., Chang, S.-C., & Yang, C.-Y. (2022). Training a Deep Contextualized Language Model for International Classification of Diseases, 10th Revision Classification via Federated Learning: Model Development and Validation Study. *JMIR Medical Informatics*, *10*(11), e41342. https://doi.org/10.2196/41342

Chen, X., & Liu, B. (2021). CRank: Reusable Word Importance Ranking for Text Adversarial Attack. *Applied Sciences*, *11*(20), Article 20. https://doi.org/10.3390/app11209570

Clement, T., Kemmerzell, N., Abdelaal, M., & Amberg, M. (2023). XAIR: A Systematic Metareview of Explainable AI (XAI) Aligned to the Software Development Process. *Machine Learning and Knowledge Extraction*, *5*(1), Article 1. https://doi.org/10.3390/make5010006

Cora Garcia, A., David, G. C., & Chand, D. (2010). Understanding the work of medical transcriptionists in the production of medical records. *Health Informatics Journal*, *16*(2), 87–100. https://doi.org/10.1177/1460458210361936

Das, B., & Chakraborty, S. (2018). *An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation* (arXiv:1806.06407). arXiv. https://doi.org/10.48550/arXiv.1806.06407

Deng, Y., Groll, M. J., & Denecke, K. (2015). Rule-based Cervical Spine Defect Classification Using Medical Narratives. In *MEDINFO 2015: eHealth-enabled Health* (pp. 1038–1038). IOS Press. https://doi.org/10.3233/978-1-61499-564-7-1038

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Di Nunzio, G. M., & Vezzani, F. (2018). A Linguistic Failure Analysis of Classification of Medical Publications: A Study on Stemming vs Lemmatization. In E. Cabrio, A. Mazzei, & F. Tamburini (Eds.), *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018: 10-12 December 2018, Torino* (pp. 182–186). Accademia University Press. https://doi.org/10.4000/books.aaccademia.3327

Eftekhari, H. (2024). Transcribing in the digital age: Qualitative research practice utilizing intelligent speech recognition technology. *European Journal of Cardiovascular Nursing*, zvae013. https://doi.org/10.1093/eurjcn/zvae013

Elhaddad, M., & Hamam, S. (n.d.). AI-Driven Clinical Decision Support Systems: An Ongoing Pursuit of Potential. *Cureus*, *16*(4), e57728. https://doi.org/10.7759/cureus.57728

Ellis, R. J., Sander, R. M., & Limon, A. (2022). Twelve key challenges in medical machine learning and solutions. *Intelligence-Based Medicine*, *6*, 100068. https://doi.org/10.1016/j.ibmed.2022.100068

Elreedy, D., & Atiya, A. F. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences*, *505*, 32–64. https://doi.org/10.1016/j.ins.2019.07.070

Fehr, J., Citro, B., Malpani, R., Lippert, C., & Madai, V. I. (2024). A trustworthy AI reality-check: The lack of transparency of artificial intelligence products in healthcare. *Frontiers in Digital Health*, *6*, 1267290. https://doi.org/10.3389/fdgth.2024.1267290

Feigl, M., Roesky, B., Herrnegger, M., Schulz, K., & Hayashi, M. (2022). Learning from mistakes—Assessing the performance and uncertainty in process-based models. *Hydrological Processes*, *36*(2), e14515. https://doi.org/10.1002/hyp.14515

Filipp, F. V. (2019). Opportunities for Artificial Intelligence in Advancing Precision Medicine. *Current Genetic Medicine Reports*, *7*(4), 208–213. https://doi.org/10.1007/s40142-019-00177-4

Frasca, M., La Torre, D., Pravettoni, G., & Cutica, I. (2024). Explainable and interpretable artificial intelligence in medicine: A systematic bibliometric review. *Discover Artificial Intelligence*, *4*(1), 15. https://doi.org/10.1007/s44163-024-00114-7

Goodwin, N. L., Nilsson, S. R. O., Choong, J. J., & Golden, S. A. (2022). Toward the explainability, transparency, and universality of machine learning for behavioral classification in neuroscience. *Current Opinion in Neurobiology*, *73*, 102544. https://doi.org/10.1016/j.conb.2022.102544

Guo, Y., Chung, F., & Li, G. (2016). An ensemble embedded feature selection method for multi-label clinical text classification. *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 823–826. https://doi.org/10.1109/BIBM.2016.7822631

Health, C. for D. and R. (2024). Artificial Intelligence and Machine Learning in Software as a Medical Device. *FDA*. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

Heimerl, F., & Gleicher, M. (2018). Interactive Analysis of Word Vector Embeddings. *Computer Graphics Forum*, *37*(3), 253–265. https://doi.org/10.1111/cgf.13417

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2022). Explainable AI Methods—A Brief Overview. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (pp. 13–38). Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_2

Hu, Y., Yu, Z., Cheng, X., Luo, Y., & Wen, C. (2020). A bibliometric analysis and visualization of medical data mining research. *Medicine*, *99*(22), e20338. https://doi.org/10.1097/MD.0000000000020338

Huang, K., Altosaar, J., & Ranganath, R. (2020). *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission* (arXiv:1904.05342). arXiv. https://doi.org/10.48550/arXiv.1904.05342

Huang, K., Garapati, S., & Rich, A. S. (2020). *An Interpretable End-to-end Fine-tuning Approach for Long Clinical Text* (arXiv:2011.06504). arXiv. https://doi.org/10.48550/arXiv.2011.06504

Janowski, A. (2023). Natural Language Processing Techniques for Clinical Text Analysis in Healthcare. *Journal of Advanced Analytics in Healthcare Management*, *7*(1), Article 1.

Johnson, K. B., Wei, W., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K., Zhao, J., & Snowdon, J. L. (2021). Precision Medicine, AI, and the Future of Personalized Health Care. *Clinical and Translational Science*, *14*(1), 86–93. https://doi.org/10.1111/cts.12884

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Juluru, K., Shih, H.-H., Keshava Murthy, K. N., & Elnajjar, P. (2021). Bag-of-Words Technique in Natural Language Processing: A Primer for Radiologists. *RadioGraphics*, *41*(5), 1420–1426. https://doi.org/10.1148/rg.2021210025

Jung, J., Lee, H., Jung, H., & Kim, H. (2023). Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon*, *9*(5). https://doi.org/10.1016/j.heliyon.2023.e16110

Kanda, E., Epureanu, B. I., Adachi, T., Tsuruta, Y., Kikuchi, K., Kashihara, N., Abe, M., Masakane, I., & Nitta, K. (2020). Application of explainable ensemble artificial intelligence model to categorization of hemodialysis-patient and treatment using nationwide-real-world data in Japan. *PLOS ONE*, *15*(5), e0233491. https://doi.org/10.1371/journal.pone.0233491

Khadhraoui, M., Bellaaj, H., Ammar, M. B., Hamam, H., & Jmaiel, M. (2022). Survey of BERT-Base Models for Scientific Text Classification: COVID-19 Case Study. *Applied Sciences*, *12*(6), Article 6. https://doi.org/10.3390/app12062891

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)* (arXiv:1711.11279). arXiv. https://doi.org/10.48550/arXiv.1711.11279

Kiseleva, A., Kotzinos, D., & De Hert, P. (2022). Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.879603

Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Exploratory Data Analysis. In MIT Critical Data (Ed.), *Secondary Analysis of Electronic Health Records* (pp. 185–203). Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_15

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text Classification Algorithms: A Survey. *Information*, *10*(4), Article 4. https://doi.org/10.3390/info10040150

Lambert, M., Schuster, T., Kessel, M., & Atkinson, C. (2023). Robustness Analysis of Machine Learning Models Using Domain-Specific Test Data Perturbation. In N. Moniz, Z. Vale, J. Cascalho, C. Silva, & R. Sebastião (Eds.), *Progress in Artificial Intelligence* (pp. 158–170). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-49008-8_13

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations* (arXiv:1909.11942). arXiv. https://doi.org/10.48550/arXiv.1909.11942

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Li, B., Zhao, Z., Liu, T., Wang, P., & Du, X. (2016). Weighted Neural Bag-of-n-grams Model: New Baselines for Text Classification. In Y. Matsumoto & R. Prasad (Eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 1591–1600). The COLING 2016 Organizing Committee. https://aclanthology.org/C16-1150

Li, X., Yuan, W., Peng, D., Mei, Q., & Wang, Y. (2022). When BERT meets Bilbo: A learning curve analysis of pretrained language model on disease classification. *BMC Medical Informatics and Decision Making*, *21*(9), 377. https://doi.org/10.1186/s12911-022-01829-2

Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., & Salimi-Khorshidi, G. (2020). BEHRT: Transformer for Electronic Health Records. *Scientific Reports*, *10*(1), 7155. https://doi.org/10.1038/s41598-020-62922-y

Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H., & Luo, Y. (2023). A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, *30*(2), 340–347. https://doi.org/10.1093/jamia/ocac225

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Lohmann, S., Heimerl, F., Bopp, F., Burch, M., & Ertl, T. (2015). Concentri Cloud: Word Cloud Visualization for Multiple Text Documents. *2015 19th International Conference on Information Visualisation*, 114–120. https://doi.org/10.1109/iV.2015.30

Lu, H., Ehwerhemuepha, L., & Rakovski, C. (2022). A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC Medical Research Methodology*, *22*(1), 181. https://doi.org/10.1186/s12874-022-01665-y

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Marafino, B. J., Davies, J. M., Bardach, N. S., Dean, M. L., & Dudley, R. A. (2014). N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *Journal of the American Medical Informatics Association: JAMIA*, *21*(5), 871–875. https://doi.org/10.1136/amiajnl-2014-002694

Mavrepis, P., Makridis, G., Fatouros, G., Koukos, V., Separdani, M. M., & Kyriazis, D. (2024). *XAI for All: Can Large Language Models Simplify Explainable AI?* (arXiv:2401.13110). arXiv. https://doi.org/10.48550/arXiv.2401.13110

Moazemi, S., Vahdati, S., Li, J., Kalkhoff, S., Castano, L. J. V., Dewitz, B., Bibo, R., Sabouniaghdam, P., Tootooni, M. S., Bundschuh, R. A., Lichtenberg, A., Aubin, H., & Schmid, F. (2023). Artificial intelligence for clinical decision support for monitoring patients in cardiovascular ICUs: A systematic review. *Frontiers in Medicine*, *10*. https://doi.org/10.3389/fmed.2023.1109411

MOS. (2021, July 1). *Progress of Medical Transcription Over Time*. https://www.medicaltranscriptionservicecompany.com/blog/how-medical-transcription-evolved-through-years/

Mujtaba, G., Shuib, L., Idris, N., Hoo, W. L., Raj, R. G., Khowaja, K., Shaikh, K., & Nweke, H. F. (2019). Clinical text classification research trends: Systematic literature review and open

issues. *Expert Systems with Applications*, *116*, 494–520. https://doi.org/10.1016/j.eswa.2018.09.034

Muller, B., Elazar, Y., Sagot, B., & Seddah, D. (2021). First Align, then Predict: Understanding the Cross-Lingual Ability of Multilingual BERT. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2214–2231). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.189

Munthuli, A., Pooprasert, P., Klangpornkun, N., Phienphanich, P., Onsuwan, C., Jaisin, K., Pattanaseri, K., Lortrakul, J., & Tantibundhit, C. (2023). Classification and analysis of text transcription from Thai depression assessment tasks among patients with depression. *PLOS ONE*, *18*(3), e0283095. https://doi.org/10.1371/journal.pone.0283095

Murdoch, B. (2021). Privacy and artificial intelligence: Challenges for protecting health information in a new era. *BMC Medical Ethics*, *22*(1), 122. https://doi.org/10.1186/s12910-021-00687-3

Naseem, U., & Musial, K. (2019). DICE: Deep Intelligent Contextual Embedding for Twitter Sentiment Analysis. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 953–958. https://doi.org/10.1109/ICDAR.2019.00157

Ngai, H., & Rudzicz, F. (2022). *Doctor XAvIer: Explainable Diagnosis on Physician-Patient Dialogues and XAI Evaluation* (arXiv:2204.10178). arXiv. https://doi.org/10.48550/arXiv.2204.10178

Nguyen, T. (2023, September 19). *Research Paper: Developing a Fair AI-based Healthcare Framework with Feedback Loop*. Ethical Computing. https://www.ethicalcomputing.auckland.ac.nz/developing-a-fair-ai-based-healthcare-framework-with-feedback-loop/

Ong, M.-S., Magrabi, F., & Coiera, E. (2010). Automated categorisation of clinical incident reports using statistical text classification. *Quality and Safety in Health Care*, *19*(6), e55–e55. https://doi.org/10.1136/qshc.2009.036657

Oubenali, N., Messaoud, S., Filiot, A., Lamer, A., & Andrey, P. (2022). Visualization of medical concepts represented using word embeddings: A scoping review. *BMC Medical Informatics and Decision Making*, *22*(1), 83. https://doi.org/10.1186/s12911-022-01822-9

Ozcan, I., Aydin, H., & Cetinkaya, A. (2022). Comparison of Classification Success Rates of Different Machine Learning Algorithms in the Diagnosis of Breast Cancer. *Asian Pacific Journal of Cancer Prevention*, *23*(10), 3287–3297. https://doi.org/10.31557/APJCP.2022.23.10.3287

Paaß, G., & Giesselbach, S. (2023). Pre-trained Language Models. In G. Paaß & S. Giesselbach (Eds.), *Foundation Models for Natural Language Processing: Pre-trained Language Models Integrating Media* (pp. 19–78). Springer International Publishing. https://doi.org/10.1007/978-3-031-23190-2_2

Pagad, N. S., N, P., Almuzaini, K. K., Maheshwari, M., Gangodkar, D., Shukla, P., & Alhassan, M. (2022). Clinical Text Data Categorization and Feature Extraction Using Medical-Fissure Algorithm and Neg-Seq Algorithm. *Computational Intelligence and Neuroscience*, *2022*, 5759521. https://doi.org/10.1155/2022/5759521

Pahwa, B., Taruna, S., & Kasliwal, N. (2018). Sentiment Analysis- Strategy for Text Pre-Processing. *International Journal of Computer Applications*, *180*(34), 15–18. https://doi.org/10.5120/ijca2018916865

Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., & Rohrbach, M. (2018). *Multimodal Explanations: Justifying Decisions and Pointing to the Evidence* (arXiv:1802.08129). arXiv. https://doi.org/10.48550/arXiv.1802.08129

Prabhakar, S. K., & Won, D.-O. (2021). Medical Text Classification Using Hybrid Deep Learning Models with Multihead Attention. *Computational Intelligence and Neuroscience*, *2021*(1), 9425655. https://doi.org/10.1155/2021/9425655

Qasim, R., Bangyal, W. H., Alqarni, M. A., & Ali Almazroi, A. (2022). A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification. *Journal of Healthcare Engineering*, *2022*(1), 3498123. https://doi.org/10.1155/2022/3498123

Qing, L., Linhong, W., & Xuehai, D. (2019). A Novel Neural Network-Based Method for Medical Text Classification. *Future Internet*, *11*(12), Article 12. https://doi.org/10.3390/fi11120255

Rai, A., & Borah, S. (2021). Study of Various Methods for Tokenization. In J. K. Mandal, S. Mukhopadhyay, & A. Roy (Eds.), *Applications of Internet of Things* (pp. 193–200). Springer. https://doi.org/10.1007/978-981-15-6198-6_18

Ramdhani, A., Ramdhani, M., & Amin, A. (2014). Writing a Literature Review Research Paper: A step-by-step approach. *International Journal of Basic and Applied Science*, *3*, 47–56.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *Npj Digital Medicine*, *4*(1), 1–13. https://doi.org/10.1038/s41746-021-00455-y

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rijcken, E., Kaymak, U., Scheepers, F., Mosteiro, P., Zervanou, K., & Spruit, M. (2022). Topic Modeling for Interpretable Text Classification From EHRs. *Frontiers in Big Data*, *5*. https://doi.org/10.3389/fdata.2022.846930

Rodriguez, A., Tuck, C., Dozier, M. F., Lewis, S. C., Eldridge, S., Jackson, T., Murray, A., & Weir, C. J. (2022). Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review. *Clinical Trials (London, England)*, *19*(4), 452–463. https://doi.org/10.1177/17407745221087469

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *PLOS ONE*, *16*(8), e0254937. https://doi.org/10.1371/journal.pone.0254937

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Schöffer, J. (2023). *On the Interplay of Transparency and Fairness in AI-Informed Decision-Making*. https://doi.org/10.5445/IR/1000164741

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Shang, J., Ma, T., Xiao, C., & Sun, J. (2019). *Pre-training of Graph Augmented Transformers for Medication Recommendation*. 5953–5959.

Shao, Y., Taylor, S., Marshall, N., Morioka, C., & Zeng-Treitler, Q. (2018). Clinical Text Classification with Word Embedding Features vs. Bag-of-Words Features. *2018 IEEE International Conference on Big Data (Big Data)*, 2874–2878. https://doi.org/10.1109/BigData.2018.8622345

Singhal, A., Neveditsin, N., Tanveer, H., & Mago, V. (2024). Toward Fairness, Accountability, Transparency, and Ethics in AI for Social Media and Health Care: Scoping Review. *JMIR Medical Informatics*, *12*(1), e50048. https://doi.org/10.2196/50048

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Sreekumar, Y., & Nizar Banu, P. K. (2022). Clinical Text Classification of Medical Transcriptions Based on Different Diseases. In Ch. Satyanarayana, D. Samanta, X.-Z. Gao, & R. K. Kapoor (Eds.), *High Performance Computing and Networking* (pp. 613–623). Springer. https://doi.org/10.1007/978-981-16-9885-9_50

Szczepański, M., Pawlicki, M., Kozik, R., & Choraś, M. (2021). New explainability method for BERT-based model in fake news detection. *Scientific Reports*, *11*(1), 23705. https://doi.org/10.1038/s41598-021-03100-6

Taherdoost, H. (2021). Data Collection Methods and Tools for Research; A Step-by-Step Guide to Choose Data Collection Technique for Academic and Business Research Projects. *International Journal of Academic Research in Management (IJARM)*, *10*(1), 10–38.

Talebi, S., Tong, E., Li, A., Yamin, G., Zaharchuk, G., & Mofrad, M. R. K. (2024). Exploring the performance and explainability of fine-tuned BERT models for neuroradiology protocol assignment. *BMC Medical Informatics and Decision Making*, *24*(1), 40. https://doi.org/10.1186/s12911-024-02444-z

Ting, K. M. (2010). Confusion Matrix. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 209–209). Springer US. https://doi.org/10.1007/978-0-387-30164-8_157

Tjoa, E., & Guan, C. (2021). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(11), 4793–4813. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2020.3027314

Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T., & Naganawa, S. (2024). Fairness of artificial intelligence in healthcare: Review and recommendations. *Japanese Journal of Radiology*, *42*(1), 3–15. https://doi.org/10.1007/s11604-023-01474-3

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). *Attention is All you Need*. Neural Information Processing Systems. https://www.semanticscholar.org/paper/Attention-is-All-you-Need-Vaswani-Shazeer/204e3073870fae3d05bcbc2f6a8e263d9b72e776

Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*, *11*(1), 44. https://doi.org/10.1186/s40537-024-00905-w

Wang, Y., Wang, Y., Peng, Z., Zhang, F., Zhou, L., & Yang, F. (2023). Medical text classification based on the discriminative pre-training model and prompt-tuning. *DIGITAL HEALTH*, *9*, 20552076231193213. https://doi.org/10.1177/20552076231193213

Weng, W.-H., Wagholikar, K. B., McCray, A. T., Szolovits, P., & Chueh, H. C. (2017). Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Medical Informatics and Decision Making*, *17*(1), 155. https://doi.org/10.1186/s12911-017-0556-8

White, J., & Power, S. D. (2023). k-Fold Cross-Validation Can Significantly Over-Estimate True Classification Accuracy in Common EEG-Based Passive BCI Experimental Designs: An Empirical Investigation. *Sensors (Basel, Switzerland)*, *23*(13), 6077. https://doi.org/10.3390/s23136077

Williamson, S. M., & Prybutok, V. (2024). Balancing Privacy and Progress: A Review of Privacy Challenges, Systemic Oversight, and Patient Perceptions in AI-Driven Healthcare. *Applied Sciences*, *14*(2), Article 2. https://doi.org/10.3390/app14020675

Win, L. K., & Hoon, G. K. (2022). Text Classification of Medical Transcriptions using N-Gram Machine Learning Approach. *2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, 1–6. https://doi.org/10.1109/IICAIET55139.2022.9936867

Yang, W., Wei, Y., Wei, H., Chen, Y., Huang, G., Li, X., Li, R., Yao, N., Wang, X., Gu, X., Amin, M. B., & Kang, B. (2023). Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. *Human-Centric Intelligent Systems*, *3*(3), 161–188. https://doi.org/10.1007/s44230-023-00038-y

Yao, J., Xu, W., Lian, J., Wang, X., Yi, X., & Xie, X. (2023). *Knowledge Plugins: Enhancing Large Language Models for Domain-Specific Recommendations* (arXiv:2311.10779). arXiv. https://doi.org/10.48550/arXiv.2311.10779

Yao, L., Mao, C., & Luo, Y. (2019). Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics and Decision Making*, *19*(3), 71. https://doi.org/10.1186/s12911-019-0781-4

Zhang, H., & Ogasawara, K. (2023). Grad-CAM-Based Explainable Artificial Intelligence Related to Medical Text Processing. *Bioengineering*, *10*(9), 1070. https://doi.org/10.3390/bioengineering10091070

Zhang, J., Wu, J., Qiu, Y., Song, A., Li, W., Li, X., & Liu, Y. (2023). Intelligent speech technologies for transcription, disease diagnosis, and medical equipment interactive control in smart hospitals: A review. *Computers in Biology and Medicine*, *153*, 106517. https://doi.org/10.1016/j.compbiomed.2022.106517

Zhao, Z., Zhang, Z., & Hopfgartner, F. (2021). A Comparative Study of Using Pre-trained Language Models for Toxic Comment Classification. *Companion Proceedings of the Web Conference 2021*, 500–507. https://doi.org/10.1145/3442442.3452313

Zhou, Z., Hu, M., Salcedo, M., Gravel, N., Yeung, W., Venkat, A., Guo, D., Zhang, J., Kannan, N., & Li, S. (2023). *XAI meets Biology: A Comprehensive Review of Explainable AI in Bioinformatics Applications* (arXiv:2312.06082). arXiv. https://doi.org/10.48550/arXiv.2312.06082

Zhu, Y., Moniz, J. R. A., Bhargava, S., Lu, J., Piraviperumal, D., Li, S., Zhang, Y., Yu, H., & Tseng, B.-H. (2024). *Can Large Language Models Understand Context?* (arXiv:2402.00858). arXiv. https://doi.org/10.48550/arXiv.2402.00858

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Declaration of Authenticity

I hereby declare that I have completed this Bachelors/ Master's thesis on my own and without any additional external assistance. I have made use of only those sources and aids specified and I have listed all the sources from which I have extracted text and content. This thesis or parts thereof have never been presented to another examination board. I agree to a plagiarism check of my thesis via a plagiarism detection service.

Berlin, 06th August 2024

Place, Date

Student signature