NLP and Deep Learning for Text and Audio Classification in Medical Diagnosis

Dissertation Manuscript

Submitted to National University

School of Technology and Engineering

In Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY
In Data Science

by

HAMEEM M MAHDI

San Diego, California

October 2025

Abstract

Medical diagnosis accuracy remains a critical challenge in healthcare, with misdiagnosis affecting millions of patients annually and contributing to adverse outcomes and increased healthcare costs. This study addresses the limited availability of computational tools for analyzing patient-generated symptom descriptions in both textual and audio formats. The research develops and evaluates a framework for automated symptom classification using natural language processing and deep learning techniques. This multimodal approach mirrors real-world clinical scenarios where practitioners gather information through various communication channels.

The study utilized publicly available medical symptom datasets containing text transcriptions and audio recordings (Medical Speech, Transcription, and Intent dataset; MIT License; available at Kaggle. A train-validation-test split methodology (70-15-15) was implemented, and six classification algorithms were evaluated across three modalities (audio-only, text-only, and audio-text fusion). Data preparation included:

- **Text modality**: Text normalization, cleaning, and feature extraction using TF-IDF vectorization (max 128 features).
- **Audio modality**: Audio preprocessing with MFCC feature extraction (13 coefficients) and feature scaling using StandardScaler.
- **Audio-text fusion**: Concatenation of normalized audio features (MFCC) and text features (TF-IDF) into unified feature vectors.

The six classification algorithms evaluated were Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), Feedforward Neural Network (FNN), and Convolutional Neural Network (CNN).

Results demonstrated varying performance across the three modalities:

- Text-only classification: Logistic Regression achieved the highest test performance, with a weighted F1-score of 91.72%, accuracy of 91.79%, precision of 92.16%, and recall of 91.79%, successfully meeting clinical deployment thresholds (>75%).
- Audio-only classification: Logistic Regression achieved the highest test performance, with a weighted F1-score of 7.33%, accuracy of 8.13%, precision of 8.45%, and recall of 8.13%, but still significantly below clinical deployment thresholds.
- Audio-text fusion classification: CNN (deep learning) achieved the highest test performance, with a weighted F1-score of 83.85%, accuracy of 85.07%, precision of 86.62%, and recall of 85.07%, successfully exceeding clinical deployment thresholds.

A train-validation-test split methodology (70%-15%-15%) confirmed consistent performance across diverse symptom categories. The complete implementation includes three separate Jupyter notebooks (text_medical_diagnosis.ipynb, audio_medical_diagnosis.ipynb, and multimodal_medical_diagnosis.ipynb) with full analysis pipelines. The complete implementation code and analysis are available at https://github.com/HAMEEMM/multimodal_medical_diagnosis.

This research develops an integrated multimodal system combining text and audio analysis to advance clinical diagnostic AI, demonstrating modality-specific algorithm performance. The approach augments clinical decision-making while preserving physician judgment, with future directions addressing ambiguous cases and multilingual coverage.

Acknowledgments

I would like to express my profound gratitude to my dissertation committee chair, Dr. Nabeel, whose guidance, expertise, and unwavering support were instrumental in shaping this research. I want to extend my sincere appreciation to my committee members, Dr. Tsapara and Dr. Al-Najada, for their valuable insights, constructive feedback, and encouragement throughout this academic journey.

I am deeply grateful to the Department of Engineering, Computer Science, and Data Science, as well as the National University, for providing the resources and academic environment that enabled me to pursue this research. Special thanks to Kaggle.com for providing access to computational resources that enabled the intensive data analysis required for this study. This research would not have been possible without the availability of open-source medical datasets. I acknowledge the creators and contributors of the Medical Speech, Transcription, and Intent dataset, whose work provided the foundation for this study.

I would like to express my deepest gratitude to my family, who have been my pillar of strength throughout this journey. To my father, whose wisdom and encouragement kept me focused on my goals; to my mother, whose unconditional love and support sustained me through challenging times; and to my brothers and sisters, whose cheerful spirit and belief in my abilities never wavered. Your collective faith in me has been my greatest source of motivation and resilience. I extend my heartfelt gratitude to my son Tim; your belief in me has been my greatest motivation. This accomplishment would not have been possible without your constant encouragement and understanding during the countless hours dedicated to this research.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Technological advancements are ushering in a transformative era in the healthcare sector, with Natural Language Processing (NLP) and Deep Learning (DL) at the forefront. NLP has the potential to significantly transform how doctors diagnose illnesses, prescribe treatments, and communicate with patients. There have also been improvements in the healthcare sector documentation through the adoption of Electronic Health Records and digital imaging solutions, which have provided the data needed to improve medical outcomes and streamline workflows. This research study will focus on Healthcare NLP and DL.

Medical text and audio classification can enhance medical treatment and diagnosis applications, thereby reducing morbidity and mortality. Kobritz et al. (2023) suggest that "Machine-learning algorithms show promise in improving prediction of complications."

This study examines the application of deep learning and natural language processing to healthcare text and audio data for disease classification. This chapter provides an overview of the clinical applications of NLP text and audio classification in treatment and diagnosis, highlighting their importance in the medical sector. The problem statement addresses challenges in medical diagnosis and treatment due to the lack of reliable tools for analyzing textual and auditory data in healthcare, aiming to improve diagnosis and enhance the effectiveness of patient treatment. The study addresses critical issues and provides a comprehensive roadmap for implementing the entire project, including, but not limited to, background, problem, purpose, variables, population, sample, and conceptual framework for this research. This study also develops research hypotheses, questions, and significance concerning the research topic. The healthcare sector is increasingly relying on cutting-edge technology to enhance patient care, streamline clinical processes, and improve diagnostic accuracy.

NLP provides essential tools for this setting (Johri et al., 2021). NLP is a set of methods for processing unstructured text. Studies indicate that implementing NLP in healthcare may improve medical diagnosis (Al-Garadi et al., 2022) because health systems are increasingly able to interpret, analyze, and search large volumes of patient information. Wang et al. (2024) found that "Recent years have witnessed a substantial increase in the use of deep learning to solve various natural language processing (NLP) problems" for many healthcare institutions. Decision-making capabilities in the medical sector have been significantly enhanced through the incorporation of NLP methods. This study will use aspect mining and sentiment analysis to extract features that yield desirable results for the algorithm during prediction.

NLP algorithms are the only ML algorithms that can potentially reduce mortality in healthcare, as patient history data can be leveraged to inform decision-making. Fagherazzi et al. (2021) argue that to maintain control over the recorded vocal task and allow patients to choose their own words, thereby preserving their naturalness, semi-spontaneous voice tasks are designed. In these tasks, the patient is instructed to discuss a specific topic (e.g., picture description or story narration).

To establish the significance of this research, it must be contextualized within the current healthcare informatics landscape, where natural language processing applications are transforming diagnostic methodologies. The research study focused on text and audio classification for medical diagnosis, an area gaining significant importance and current relevance. For example, applying NLP and DL to medical diagnosis has significant practical implications, including, but not limited to, using patient data to train DL models to predict disease progression and potential patient outcomes. Therefore, this information can significantly assist healthcare providers in developing effective treatment plans and enhancing patient care. NLP can also save medical personnel time and burdensome administrative work by automatically extracting relevant information from clinical notes and creating

documentation in a simplified format. The ability to accelerate and improve diagnostic accuracy in healthcare has a substantial impact on patient care and outcomes.

This research quantitatively analyzes the impact of NLP and DL on healthcare diagnostic procedures. The study employs these computational approaches on an existing dataset to develop classification models, thereby establishing the current state of the field and providing a foundation for the proposed research. Advancements and limitations identified in the literature review underscore the significance of addressing the research problem.

1.1 Statement of the Problem

1.1.1 Core Problem

The fundamental challenge this study addresses is the underutilization of textual and auditory data in clinical decision-making (Lu et al., 2020). Despite the abundance of patient-generated data in modern healthcare settings, the absence of systematic analytical frameworks impedes its integration into diagnostic workflows. Current healthcare systems rely predominantly on subjective human interpretation of complex medical narratives and auditory cues, which introduces significant variability in patient assessment and treatment planning.

1.1.2 Consequences

This deficiency manifests in multiple adverse outcomes across the healthcare ecosystem. For clinicians, the cognitive burden of processing unstructured data contributes to decision fatigue and increases the likelihood of diagnostic errors (Stark et al., 2018). Patients experience delayed diagnoses, suboptimal treatment regimens, and diminished health outcomes. At the organizational level, healthcare institutions face escalating operational costs, increased litigation risk, and compromised quality metrics. Heys et al. (2022) highlight that legacy documentation systems—primarily retrospective and paper-based—further exacerbate these issues by creating informational silos that prevent timely data retrieval and analysis.

1.1.3 Research Gap

While previous studies have explored computational approaches to medical data analysis, a significant gap exists in developing integrated frameworks that simultaneously process textual and auditory clinical data with sufficient accuracy and interpretability for practical implementation. The interrelationships between stakeholder needs, technical limitations, and organizational constraints remain inadequately conceptualized in the literature, limiting the translation of computational advances into clinical practice.

1.1.4 Proposed Approach

This research addresses these deficiencies by developing and validating specialized Natural Language Processing (NLP) and Deep Learning (DL) models tailored to the unique characteristics of healthcare data. This study aims to enhance diagnostic precision, improve treatment selection, and optimize resource allocation across healthcare systems by creating analytical frameworks that transform unstructured clinical information into actionable insights. Recent findings by Jain et al. (2023) demonstrate that machine learning and deep learning models achieve accuracies exceeding 90%, providing a promising foundation for this approach.

1.2 Purpose of the Study

This research aims to develop integrated computational models that enhance clinical decision-making by automating the analysis of patient-generated text and audio data. The study addresses the diagnostic inefficiencies identified in the problem statement by developing analytical frameworks that transform unstructured clinical information into actionable insights.

1.2.1 Primary Objectives

The study pursues three interconnected objectives:

1. To design and validate natural language processing algorithms capable of extracting clinically relevant features from patient-reported textual descriptions of symptoms
2. To develop deep learning architectures that effectively classify auditory data containing diagnostic indicators not captured in textual records
3. To integrate these computational approaches into a unified multi-modal framework that synthesizes both information streams to improve diagnostic accuracy

1.2.2 Scope and Significance

This research focuses specifically on symptom classification as the critical first step in the diagnostic pathway. The study addresses a significant gap in current clinical practice: valuable diagnostic information is underutilized because models systematically analyze patient communications across multiple modalities. The potential impact extends to multiple stakeholders—clinicians benefit from enhanced decision support, patients receive more accurate and timely diagnoses, and healthcare institutions improve resource utilization and quality metrics.

1.2.3 Target Audience and Context

The primary beneficiaries of this research include clinical practitioners responsible for initial patient assessment and diagnosis, as well as patients seeking more efficient healthcare interactions. The study utilizes anonymized, open-source datasets comprising paired audio and textual observations to ensure ethical compliance while maintaining clinical relevance.

Through this focused approach to computational medicine, the study aims to establish a methodological foundation for integrating advanced analytical techniques into routine clinical practice.

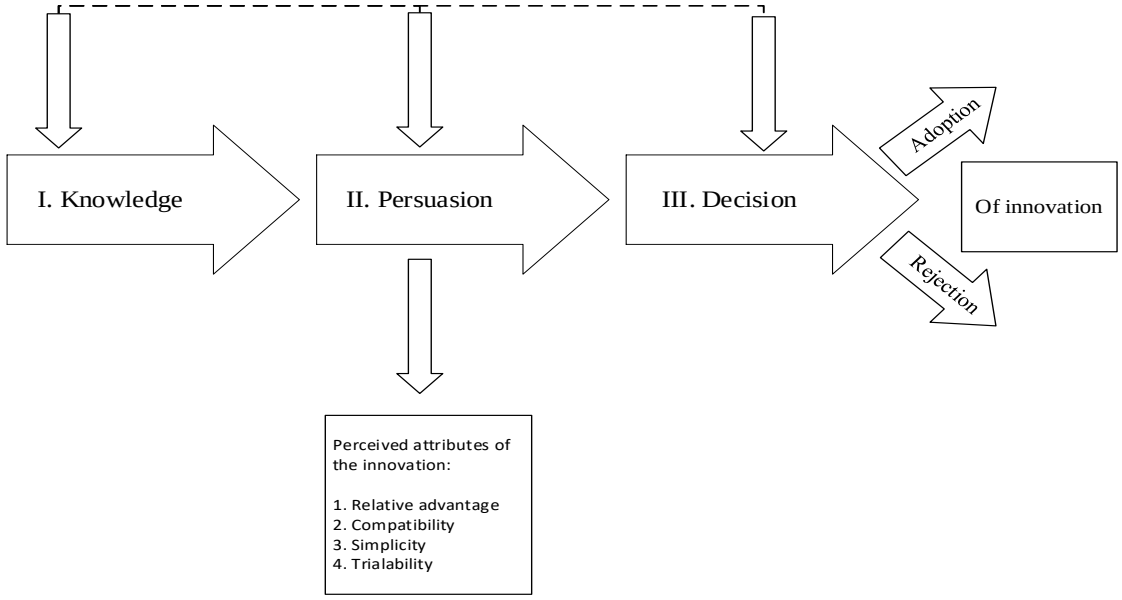
1.2.4 Introduction to the Theoretical Framework

The study examines the impact of advancements in Natural Language Processing (NLP) and Deep Learning (DL) on the healthcare industry. Therefore, this is particularly relevant. Decisions about the study's research are based on implementing the hybrid natural language processing and deep learning model. The issue statement is better shaped by considering the sector's complexity and the demand for novel approaches. The statement of purpose aligns with these frameworks, as it emphasizes the potential widespread adoption of NLP and DL in the healthcare industry, driven by the novelty of these technologies themselves (Bianchini et al., 2020). These unified models also guide research on the spread of NLP and DL innovations within the complex healthcare system.

According to Pearson et al. (2023), "Everett Rogers' diffusion theory of innovations has been widely applied to examine the timeline for the adoption of new ideas in several domains ranging from agriculture to healthcare, but has not been used to better understand the adoption of new medications, technologies, and procedures in anesthesia practice." Combining these theoretical frameworks provides a more comprehensive understanding of the diffusion of NLP and DL advances in healthcare. This research technique ensures that research decisions are grounded in theory and evidence. The integrated approach provides a more comprehensive understanding of the advantages and disadvantages of implementing cutting-edge technologies in healthcare practices.

Figure 1

The Process of Adoption in Rogers's Diffusion of Innovation Model



Note. The figure Is drive in "Evaluating the Adoption of Evidence-based Practice using Rogers's Diffusion of Innovation Theory: A Model Testing Study" by Mohammadi et al ., 2018.

This research applies Rogers' Diffusion of Innovation framework to conceptualize the implementation challenges of NLP and DL technologies in clinical settings. The five stages of Rogers' model—knowledge, persuasion, decision, implementation, and confirmation—provide a structured approach for analyzing the adoption of computational diagnostic tools by healthcare stakeholders.

This study's knowledge stage involves clinicians' awareness of NLP and DL capabilities for symptom classification. The persuasion stage addresses how evidence of diagnostic accuracy influences attitudes toward adoption. The decision, implementation, and confirmation stages map to clinical validation, workflow integration, and long-term evaluation of these technologies.

This theoretical framing enables the systematic identification of adoption barriers specific to computational diagnostics in healthcare settings. By understanding these barriers, the study can propose implementation strategies that account for the unique organizational, professional, and ethical considerations in the diffusion of medical innovation.

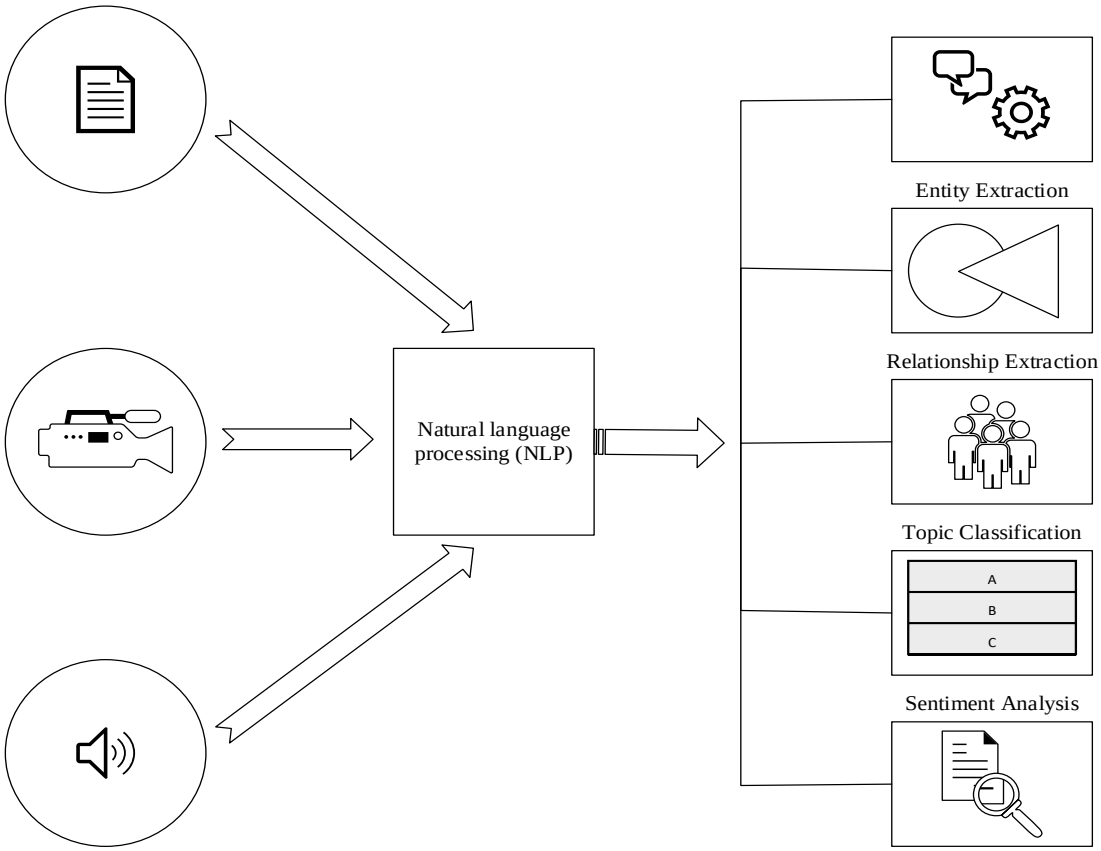
1.3 Introduction to Research Methodology and Design

This section establishes the methodological framework for investigating how NLP and DL can enhance medical diagnosis through computational analysis of patient-generated data. Rather than simply describing technologies, this research adopts a constructive approach to develop and validate practical computational solutions for clinical applications.

1.3.1 Research Design and Approach

The study employs a constructive research methodology that focuses on building and evaluating computational models to address real-world diagnostic challenges in healthcare. This approach aligns with the research objectives by enabling the systematic development and testing of NLP and DL architectures against defined performance metrics. Using open-access medical datasets from Kaggle.com provides a foundation for reproducible experimentation while ensuring ethical compliance through the use of anonymized data.

Figure 2
Natural Language Processing



Note. The figure is drive in " Artificial Intelligence Tools and Their Usage in Health Care Access and Quality" by Sai Ganesh et al., 2023

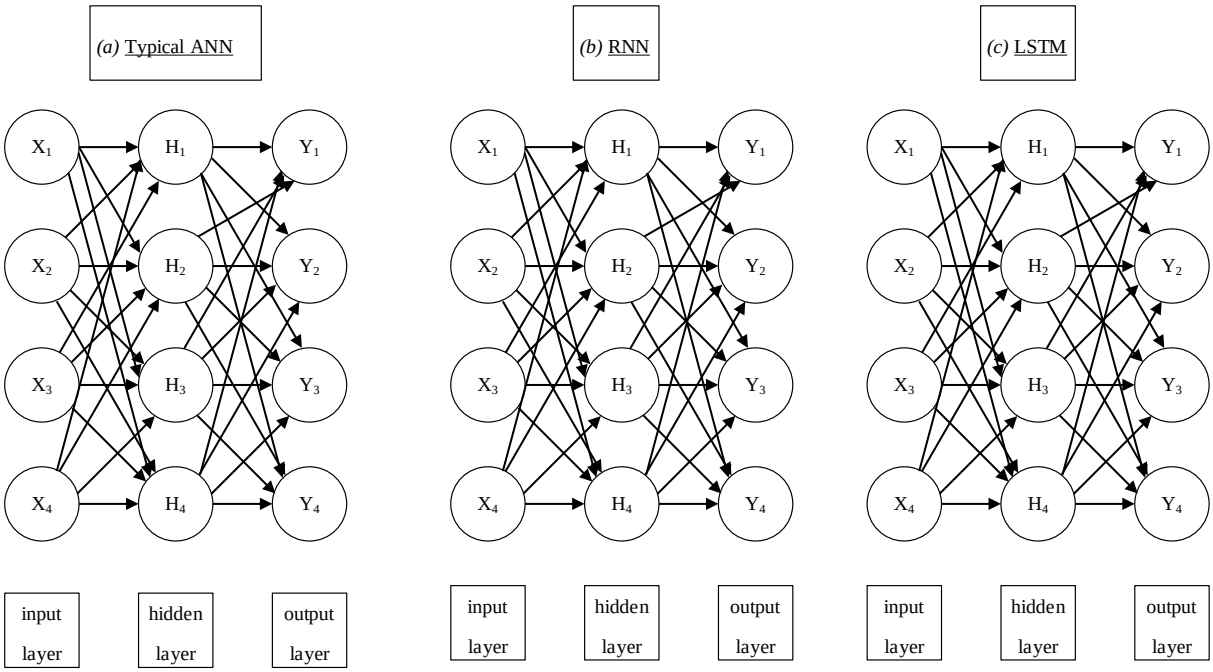
1.3.2 Computational Framework

The methodological framework integrates two complementary analytical pathways:

1. Text-Based Symptom Classification (text_medical_diagnosis.ipynb): This pathway utilizes six classification algorithms—Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), Feedforward Neural Network (FNN), and Convolutional Neural Network (CNN)—to classify medical symptoms from textual descriptions. The approach focuses on effective textual analysis through comprehensive preprocessing (text normalization, lemmatization, stopword removal) and TF-IDF vectorization (max 128 features) for feature extraction.

Figure 3

Typical examples of ANN, RNN, and LSTM

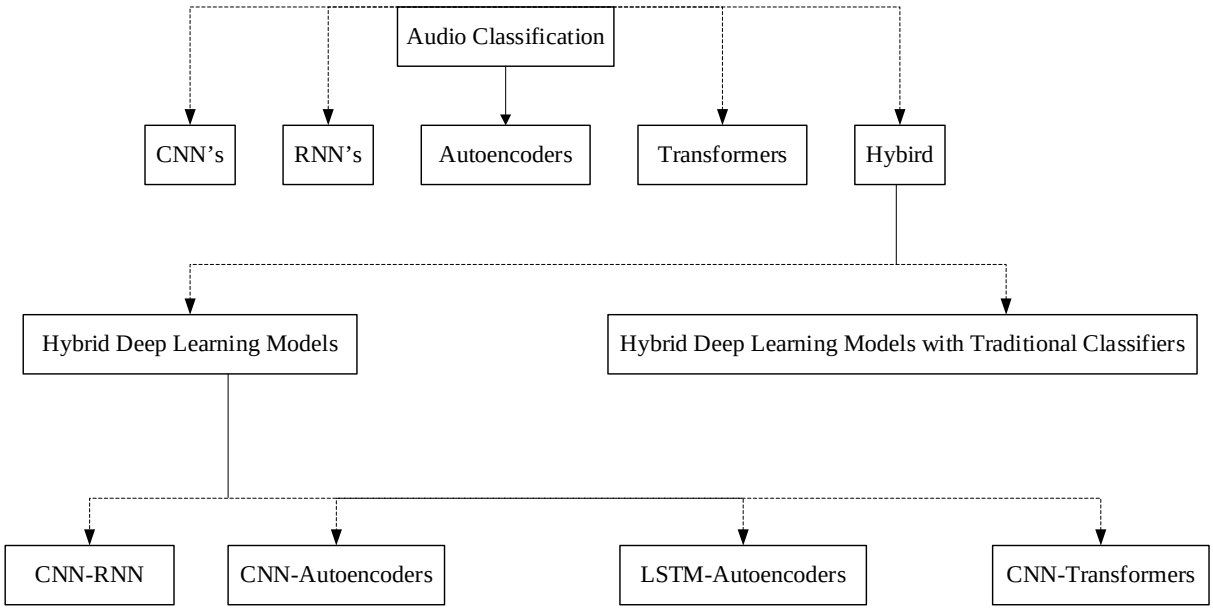


Note. The figure drive in " Spatiotemporal Prediction of PM2.5 Concentrations at Different Time Granularities Using IDW-BLSTM" by Jun et al., 2019

2. Audio-Based Symptom Classification (audio_medical_diagnosis.ipynb): This pathway analyzes patient voice recordings using acoustic feature extraction (13 MFCCs, spectral, temporal, and harmonic features—74 total features) and StandardScaler normalization. Six algorithms were evaluated using the same train-validation-test methodology.
3. Multimodal (Audio-Text) Symptom Classification (multimodal_medical_diagnosis.ipynb): This pathway integrates text and audio modalities at the feature level, concatenating TF-IDF text features with normalized MFCC audio features into unified feature vectors. The same six classification algorithms were trained on the combined feature matrix.

Figure 4

Classification of Related Work on Different Deep Learning Architectures



Note. The figure drive in " A Survey of Audio Classification Using Deep Learning by Zaman et al ., 2023.

1.3.3 Analytical Procedures

The methodological workflow comprises four sequential phases:

1. Data Preparation: Systematic cleaning and normalization procedures remove inconsistencies, duplicate entries, and null values that could compromise model performance.
2. Feature Engineering: For textual data, NLP techniques, including sentiment analysis and aspect mining, extract clinically relevant features, as validated by Al-Razgan et al. (2021). Figure 2 illustrates how diverse inputs are transformed through the NLP pipeline into structured analytical outputs.
3. Model Development: Multi-modal architectures integrate textual and audio processing streams to produce unified diagnostic predictions.

4. Performance Evaluation: Model validation employs multiple complementary metrics—accuracy, precision, F1-score, and confusion matrices—to comprehensively assess classification performance.

This methodological approach enables rigorous evaluation of the hypothesis that integrated computational analysis of patient-generated data can enhance diagnostic accuracy in clinical settings.

1.4 Research Questions

Research Questions (RQs) are the guiding queries that frame the inquiry and should be crafted to align precisely with the problem statement and the study's objective.

1.4.1 RQ1

How effective is the NLP algorithm in classifying patient symptoms from text data on the population level?

1.4.2 RQ2

How effective is the NLP algorithm in classifying patient symptoms from audio data at the population level?

1.4.3 RQ3

How effective is NLP in classifying patient symptoms from combining audio and text data on the population level?

1.5 Hypotheses

Based on the research objectives, the hypothesis for this research is as follows:

1.5.1 $H1_0$

Text analysis of patient symptoms results in insufficient precision and recall for provider decision support.

1.5.2 $H1_a$

Text analysis of patient symptoms results in precision and recall sufficient for provider decision support.

1.5.3 $H2_0$

Audio analysis of patient symptoms yields both precision and recall metrics that are insufficient for effective provider decision support.

1.5.4 $H2_a$

Audio analysis of patient symptoms results in precision and recall sufficient for provider decision support.

1.5.5 $H3_0$

Audio and text analysis of patient symptoms yields both precision and recall metrics that are insufficient for effective provider decision support.

1.5.6 $H3_a$

Audio and text analysis of patient symptoms results in precision and recall sufficient for provider decision support.

1.6 Significance of the Study

The study will demonstrate how NLP and DL might provide decision support to providers. Given the problem statement, it has been identified that the presence of unstructured medical data has made it more difficult for physicians to document clinical notes properly

(Deshmukh, 2023). As a result, a gap exists in ethical perspectives regarding the meeting of patients' expectations for accurate clinical decisions.

This research aims to demonstrate the potential of natural language processing and deep learning to enhance diagnostic accuracy through computational analysis of patient-generated data. This study holds value for both applied and academic fields. First and foremost, this research addresses a significant issue in healthcare by leveraging NLP and DL to improve symptom classification, ultimately leading to better medical diagnosis and treatment. Patient care might be significantly enhanced if the findings of this study were implemented. Diagnostic errors would be reduced, and treatment decisions would be made more quickly (Wu et al., 2020).

In addition, the results of this research contribute to both practical and theoretical dimensions of computational medicine. From a theoretical perspective, this work extends Rogers' Diffusion of Innovation framework by identifying the barriers to the adoption of NLP technologies in clinical settings. The findings refine our understanding of how computational approaches integrate with existing diagnostic paradigms, particularly addressing the theoretical tension between algorithmic prediction and clinical judgment. By demonstrating the performance characteristics of multi-modal analysis systems, this research advances conceptual models of how linguistic and auditory data can be computationally synthesized to improve diagnostic reasoning. This research contributes to the academic discussion on the incorporation of innovation in healthcare by grounding it in the PhD's Theory of Diffusion of Innovations and the Field Theory of Health Services (Fahy et al., 2020). The study's findings will provide essential insights into the theoretical underpinnings of innovation dissemination in a complex healthcare environment. In conclusion, this research is noteworthy for its potential impact on healthcare, its increased understanding of NLP and DL's practical applications, and its development of theoretical frameworks.

1.7 Definition of Key Terms

1. **Artificial Intelligence (AI)**

Artificial Intelligence **(**AI) is a discipline that explicitly develops computer models to perform tasks that require essential human intelligence.

2. **Deep Learning (DL)**

Deep Learning is a machine learning and artificial intelligence approach that aims to model and solve complex problems (Castiglioni et al., 2021). This classifier will be deployed to improve the accuracy of medical diagnosis and treatment by using deep neural networks to interpret and process audio data.

3. **Convolutional Neural Networks (CNNs)**

Convolutional neural networks (CNNs) are a subset of AI that have emerged as powerful tools for tasks such as image recognition, speech recognition, natural language processing (NLP), and, in the field of genomics, DNA sequence classification (Krichen, 2023).

4. **Feedforward Neural Networks (FNNs)**

Feedforward Neural Networks (FNNs) are machine learning algorithms modeled after the structure and function of the human brain.

5. **Machine Learning**

A subset of artificial intelligence that enables systems to learn and improve from data without being explicitly programmed, by identifying patterns and making predictions through algorithmic training. It encompasses supervised learning (labeled data), unsupervised learning (unlabeled data), and reinforcement learning approaches to solve classification, regression, and clustering problems.

6. **Logistic Regression**

A supervised learning algorithm used for binary and multiclass classification that estimates the probability of an instance belonging to a particular class using the logistic (sigmoid) function. It produces probability outputs between 0 and 1 by finding the optimal linear decision boundary, making it interpretable and computationally efficient for clinical classification tasks.

7. **Random Forest**

An ensemble learning algorithm that constructs multiple decision trees from random subsets of data and features, then combines their predictions through majority voting (classification) or averaging (regression) to improve accuracy and reduce overfitting. It provides feature importance rankings and handles both linear and non-linear relationships, making it robust for diverse medical classification tasks.

8. **Support Vector Machine (SVM)**

A supervised learning algorithm that finds the optimal hyperplane to maximize the margin between different classes, effectively separating data points with maximum distance from the decision boundary. It supports both linear and non-linear classification via kernel functions (linear, RBF, polynomial), making it well-suited to high-dimensional clinical data and demonstrating superior recall in audio-based medical symptom classification.

9. **Naïve Bayes**

Naïve Bayes classification classifies test data by learning the ground truth from clustering. Physical engineering analysis is used as a feature selection method prior to clustering (Efendy et al., 2023).

10. **Natural Language Processing (NLP)**

The study of how computers and humans communicate, known as Natural Language Processing, falls under the umbrella of Artificial Intelligence (AI) (GOYAL, 2023). It involves creating algorithms and models that enable computers to understand, interpret, and generate human language. In the context of this research, NLP refers to the use of these methods to analyze and act upon textual medical records for diagnosis and therapy.

11. Classification

A supervised learning task that assigns input instances to predefined discrete categories or classes based on patterns learned from labeled training data. In medical diagnosis, classification algorithms predict symptom categories or disease labels by learning decision boundaries that distinguish between different diagnostic outcomes from textual, audio, or multimodal patient data. Text classification has evolved due to this challenge. It is defined as assigning text documents to one or more categories (called labels) based on their content and semantics (Dogra et al., 2022). Zaman et al (2023) shared that deep learning can be used for audio signal classification in various ways.

12. Confusion Matrix

A performance evaluation tool for classification models that tabulates true positives, true negatives, false positives, and false negatives to visualize prediction accuracy across each class. It enables the calculation of diagnostic metrics, including precision, recall, F1-score, and sensitivity/specificity, providing detailed insight into model performance across all symptom categories in medical diagnosis tasks.

13. Hypothesis (H)

H is a symbol used to denote a hypothesis —an initial assumption or proposed explanation that guides experimental design and model development —and serves as a testable prediction about relationships between variables or model performance. In machine learning research, a

hypothesis is the learned function or predictive model that maps input features to output predictions, which is then validated on test data to assess statistical significance and generalization capability.

1.8 Summary

This study addresses the critical gap in clinical decision support by investigating computational approaches to medical data analysis. Through a quantitative research design, it develops and evaluates integrated natural language processing and deep learning models that analyze both textual and auditory patient data to enhance diagnostic accuracy. The significance of this work extends beyond technical performance metrics to examine the broader challenges of adoption within healthcare systems. Within Rogers' Diffusion of Innovation framework, the research explores how computational diagnostic tools progress from knowledge acquisition to implementation across the adoption stages. This study comprehensively assesses how these technologies may transform clinical practice by evaluating technical performance metrics (precision, recall, F1-score) and adoption variables (perceived advantage, compatibility, complexity). This dual focus on technical validation and implementation science addresses a significant gap in the literature, where technological innovations often fail to translate into sustained clinical adoption despite promising performance characteristics.

Chapter 2: Literature Review

2.1 Diagnostic Challenges in Contemporary Healthcare

Healthcare systems face significant challenges in diagnostic accuracy and efficiency despite technological advancements. Lu et al. (2020) identified the underutilization of multimodal patient data—mainly textual and auditory information—as a critical gap in current practice. Contemporary diagnostic processes remain heavily dependent on human interpretation of medical records and patient narratives, introducing variability that can compromise diagnostic precision (Heys et al., 2022). This variability manifests across healthcare delivery systems as diagnostic delays, treatment errors, and missed opportunities for early intervention.

The magnitude of diagnostic error presents a substantial burden to healthcare systems. Newman-Toker et al. (2020) documented that diagnostic inaccuracies contribute to approximately 10% of patient mortality and may be implicated in up to 17% of adverse events. Johns Hopkins patient safety researchers estimated that diagnostic errors account for approximately 250,000 deaths annually in the United States alone, exceeding mortality from respiratory diseases (Kim & Lee, 2021). These statistics underscore the urgency of addressing diagnostic limitations through systematic innovations.

2.2 Computational Approaches to Clinical Decision Support

The emergence of natural language processing and deep learning technologies offers promising avenues for enhancing diagnostic processes. These computational approaches allow for systematic analysis of unstructured clinical data that has historically been underutilized in diagnostic decision-making. Alturaiki et al. (2023) demonstrated that NLP techniques can

extract clinically relevant information from narrative text with increasing accuracy, potentially augmenting physician judgment in complex cases.

The application of computational methods to clinical diagnosis represents a shift from traditional models of medical decision-making. Leary et al. (2021) argued that effective diagnosis should remain central to healthcare delivery regardless of technological sophistication. This perspective emphasizes that computational approaches should complement rather than replace clinical expertise—a consideration that directly informs the design of decision support systems.

2.3 Economic and Psychosocial Implications

Beyond clinical outcomes, diagnostic inaccuracies generate substantial economic and psychosocial consequences. Financial analyses indicate that diagnostic errors impose costs exceeding one billion dollars annually on the U.S. healthcare system through unnecessary testing, inappropriate treatments, and preventable hospitalizations. These expenditures represent financial waste and opportunity costs that affect healthcare delivery across populations.

The psychosocial dimensions of diagnostic error extend beyond economic considerations. Patients who experience false-positive or false-negative diagnoses often report heightened anxiety, diminished trust in healthcare providers, and reduced adherence to treatment recommendations. These factors create a cascading effect in which initial diagnostic errors can compromise future healthcare interactions and outcomes.

2.4 Technological Innovation and Adoption

Integrating computational approaches into clinical practice requires considering implementation challenges beyond technical performance. The COVID-19 pandemic highlighted limitations in diagnostic technologies and their deployment, demonstrating that

technical capacity alone is insufficient without appropriate implementation frameworks (Kim & Lee, 2021). A systematic analysis of historical data from infectious disease outbreaks and non-communicable disease management reveals consistent patterns in technology adoption barriers that transcend specific clinical contexts.

Emerging research suggests that multi-modal analytical approaches—combining textual, auditory, and other clinical data streams—may offer superior diagnostic support compared to single-modality systems. These integrated approaches align with the complexity of clinical reasoning, which naturally synthesizes diverse information sources in formulating diagnostic hypotheses. The development of such systems represents a promising direction for enhancing diagnostic accuracy while maintaining alignment with clinical workflow requirements.

2.5 Research Gap and Opportunity

Despite promising advances in computational approaches to clinical decision support, significant gaps remain in understanding how these technologies perform across diverse patient populations and clinical contexts. The literature reveals limited investigation of integrated text-audio analysis systems designed explicitly for symptom classification and diagnostic support. Additionally, systematic evaluation of these technologies within Rogers' diffusion of innovation framework remains underexplored, creating uncertainty about implementation pathways for promising computational approaches.

This research addresses these gaps by developing and evaluating a multi-modal computational system for symptom classification while examining adoption factors that influence clinical implementation. Through this dual focus, the study contributes to the technical advancement of computational medicine and the theoretical understanding of technology diffusion in healthcare settings.

Poor diagnostic prevalence is caused by factors such as inadequate preparation and professional knowledge, hospital system problems, and technological limitations. Scholars have identified that tool diagnostic errors occur in 5% of all adult primary care visits, indicating the tremendous scope of this problem, and this requires succinct and appropriate problem-solving skills to ensure that rapid plans, which enhance implementation, have been effectively achieved through proper diagnosis.

A concerted effort is needed at various levels to prevent the consequences of inaccurate diagnosis: regular training of healthcare professionals and the implementation of modern diagnostic systems with operational techniques to prevent miscalculations. With significant investments in standardizing diagnostic processes, healthcare organizations can enhance patient outcomes while reducing costs and maintaining positive provider-patient relationships.

Literature Review: refers to the critical appraisal of other people's work and identifying how much they have achieved, what gaps in their research can be harnessed from them, and what features need to be addressed to ensure that they thoroughly address the incomplete challenges in their research. This research discusses the conceptual Framework to identify how the factors are interrelated, leading to poor disease diagnosis and deaths. The literature review will be organized into the following sections.

Introductions: This section will include the following subheadings: the problem statement and purpose of the study; an overview of the literature review; and the database, search engines, and terms used in the reports.

Theoretical Framework: This section will discuss the following domains: the description of the theoretical framework, its origin and development, existing literature reviews, and a selected framework relevant to the present-day study.

Subtopics: These subtopics include, but are not limited to, critically analyzing the entire body of reviewed articles, balancing critical review, and addressing issues of bias in the study.

Research deficiency: This section is essential because it involves a thorough analysis of different materials, which is necessary to address the deficiency in some research that needs to meet the targets effectively at all times.

Summary: This section highlights all critical areas discussed and identified in the critical review and prospective challenge, thereby defining the research gap for the application.

2.6 Databases searched in the research.

Databases are critical, as they form the source from which we conduct our research. The database streamlines the literature review process. Databases offer advanced features like search, citation tracking, and bibliography management tools. Therefore, scholars utilize various database platforms to access relevant literature, appraise other works, identify gaps, synthesize findings, and conduct informed analyses. The databases used include, but are not limited to, the following.

ACM Digital Libraries: This is a comprehensive association for computing machinery libraries, comprising records and bibliographies that contain information on computing and information technology. These libraries contained articles offering materials on natural language processing and deep learning algorithms.

PubMed: This online tool focuses on the MEDLINE databases of references and, more importantly, on biomedical topics. This was crucial, as the research focused solely on the medical fields to address the challenges posed by inappropriate and inaccurate diagnoses based on a patient’s medical history.

arXiv: This preprint repository focuses on diverse areas of computer science and other fields such as mathematics and physics. Therefore, the database may have contained relevant materials for our study and thus should have been included in the list.

Scopus: This is envisaged to be the most significant source of abstract and citation databases for all peer-reviewed literature, and it could also have been better placed to help us conduct appropriate literature reviews from the search terms. The repository may include papers on natural language processing and deep learning.

SpringerLink: This site offers access to a wide range of scientific documents, including journals, reference works, and books. It covers different disciplines, including, but not limited to, computer science. This platform may offer relevant literature for Natural language processing and Deep learning.

2.6.1 Search Engines

(i) **Google Scholar:** This powerful search engine enables us to find research papers across all academic disciplines, often providing links to full-text PDF files. This is useful as it identifies the most effective mechanism for finding other related articles.

(ii) **Core**: The core is an academic search engine dedicated to open-access research papers. The search result provided a link to relevant materials that could be used to implement systematic literature reviews.

2.6.2 Search Keys

The systematic literature review will identify critical literature, and a different database will be used to ensure that all relevant materials have been identified, thereby enhancing the selection of appropriate documents that meet the specific criteria for our research.

((" Text *" OR " Audio NLP *" OR "fatal" OR "treatment* ") AND (" technology* usage" OR " Classification Enabled " OR " disease" OR "convoluted" OR " ai " OR " artificial

intelligence " OR " machine learning " OR " natural and language and processing " OR " knowledge and engineering ") AND (" Text and Audio * " OR " Diagnosis * " OR " Natural * " OR " Language Processing " OR Deep Learning (DL) *))

The literature review spans 5 years, and the types of literature encompassed here include article reviews and meta-analyses.

2.7 Conceptual Framework of the Study

2.7.1 Theoretical Foundation: Integration of NLP and Deep Learning in Medical Diagnostics

This research is grounded in the theoretical convergence of natural language processing and deep learning as complementary computational approaches to clinical decision support. Unlike previous frameworks that treat these domains as separate technological tracks, this study adopts an integrative perspective that recognizes their synergistic potential, specifically within diagnostic contexts. The framework builds on Raghu Etukuru's (2024) conceptualization of multimodal computational analysis while addressing the limitations of the identified sequential processing.

The foundation of this framework rests on three theoretical pillars:

1. **Computational semantics**: The extraction of clinically relevant meaning from unstructured patient-generated data.
2. **Temporal pattern recognition**: Identifying time-dependent symptom manifestations across textual and auditory domains.
3. **Multi-modal integration**: The synthesis of diverse data streams into unified diagnostic representations.

Each pillar addresses specific constraints in current diagnostic practices while establishing a theoretical basis for enhanced computational approaches to symptom classification.

2.8 Critical Analysis of Model Architectures for Clinical Applications

2.8.1 Recurrent Neural Networks in Temporal Symptom Analysis

RNNs offer theoretical advantages in processing sequential medical narratives by maintaining state information across time steps. However, Swarnendu et al. (2019) demonstrated significant limitations in their application to clinical text, particularly with respect to long-term dependencies in complex medical histories. This architectural constraint manifests as diminishing performance with increasing narrative complexity—a critical limitation in medical contexts, where symptom relationships may span significant temporal distances.

2.8.2 Convolutional Approaches to Feature Extraction

CNNs excel in extracting localized patterns from both textual and auditory data, making them theoretically suitable for identifying specific symptom signatures. Their strength in feature detection is counterbalanced by fundamental limitations in handling variable-length inputs, as documented by Zhang and Shafiq (2024). This architectural constraint is particularly problematic in clinical settings where symptom descriptions vary dramatically in length and structure across patient populations.

2.8.3 Transformer-Based Architectures for Contextual Understanding

The emergence of bidirectional encoder representations (BERT) models represents a significant theoretical advancement in contextual understanding. Raza et al. (2024) demonstrated superior performance in medical classification tasks through pre-training on domain-specific corpora. However, critical analysis reveals persistent challenges in adapting

these models to specialized medical vocabularies and the computational resources required for implementation in resource-constrained clinical environments.

2.9 Research Gaps in Computational Diagnostic Support

Critical evaluation of the literature reveals three significant gaps that this study addresses:

1. **Limited multi-modal integration**: Existing research predominantly focuses on single-modality analysis (text or audio), with insufficient investigation of integrated approaches that reflect clinical reasoning processes.
2. **Inadequate validation across diverse clinical contexts**: Most studies demonstrate proof-of-concept in controlled environments but lack rigorous evaluation across varied patient populations and healthcare settings.
3. **Insufficient attention is given to implementation factors**: Technical performance metrics dominate the literature, while factors influencing clinical adoption remain underexplored, thereby limiting the translation of findings to practice.

These gaps highlight the need for a comprehensive conceptual framework that addresses technical performance and implementation considerations in computational diagnostic support.

2.10 Proposed Conceptual Model

This study proposes an integrated conceptual model synthesizing technological capabilities with implementation science. The model consists of four interrelated components:

1. **Multi-modal feature extraction**: Specialized processing pathways for textual and auditory data that preserve modality-specific information while enabling cross-modal integration
2. **Contextual representation learning**: Transformer-based architectures adapted to clinical vocabularies that capture semantic relationships unique to symptom descriptions

3. **Classification architecture**: Hybrid deep learning approaches that leverage the complementary strengths of different architectural paradigms while mitigating their limitations
4. **Implementation assessment framework**: Structured evaluation of adoption factors derived from Rogers' Diffusion of Innovation theory, specifically addressing perceived advantage, compatibility, complexity, and trialability in clinical settings

This conceptual model guides the technical development of the computational system and the evaluation of its implementation potential, addressing the identified gaps in current research while establishing a theoretical foundation for the study's methodological approach.

2.10.1 Origin of the Conceptual Framework.

NLP's disease classification paradigm has been a transformative journey for decades of invention and investigation. The history of this field can be traced back to the mid-20th century, when research on language-processing machines began (Chakraborty et al., 2024). Thanks to Alan Turing and Warren Weaver, the concepts of natural language machines that could understand and produce language like humans began to emerge.

Nonetheless, until the end of the 20th century, diagnosis was the central application of NLP, with a primary focus on condition categorization. Classifying diseases with NLP experienced a breakthrough in the 1980s, largely attributed to the emergence of expert systems and rule-based approaches (Krishnan et al., 2023). The amusing business would include the idea that these technologies involved deductive reasoning and predefined disease classification principles. They accomplish objectives, but their immobility and inability to create natural language expressions limit their effectiveness as a communication platform. The shift to statistics and machine learning is one of the hallmarks of the successful NLP revolution in the 1990s (Daigrepont, 2024). The researchers delved into in-depth scientific work, utilizing tools

such as semantic parsing, machine translation, and information retrieval to extract and store data from medical texts (Kraljevski et al., 2023). In this era, SVM (Support Vector Machine) and Naive Bayes-based algorithms paved the way for the development of the avenged disease classification system.

The 21st century has witnessed a significant revolution in NLP disease classification with the emergence of deep learning techniques (Zhang et al., 2024). With abundant data and computational power, deep neural networks have revolutionized the industry, enabling the creation of models that can automatically learn the abstract syntax and semantics of text data. This developed powerful methods of disease classification that could process extensive medical texts with greater precision.

One significant milestone in the advancement of NLP, particularly in disease classification applications, is the use of electronic health records (EHRs) for data analysis (Priyadarshini et al., 2023). The availability of healthcare data in digital format has empowered researchers to apply natural language processing (NLP) techniques to unstructured clinical records, extracting crucial information about a patient's medical history from electronic health records (EHRs) (Yagi et al., 2023). This information, including patient symptoms, diagnoses, treatment plans, and disease progression, has proven instrumental in enabling physicians to achieve unparalleled accuracy in disease classification and patient care, underscoring the importance of this development.

The application of NLP with ontologies and semantic web technologies has enabled the classification of illnesses to a higher level of granularity. These techniques enable the representation of medical facts in a formal, computer-generalizable structure, ensuring interoperability and knowledge sharing across different healthcare systems (Akhter et al., 2024).

The NLP models designed for medical domains are being explicitly developed and refined to meet the specific requirements of oncology, cardiology, and radiology. These domain-specific models refer to specialist vocabularies in diseases, helping to increase the overview and accuracy of the classification of those illnesses (Khalifa & Albadawy, 2024).

Natural language processing plays a crucial role in diagnosing, prognosing, and even treating diseases. The events currently unfolding in deep learning, which are expected to be enhanced by multimodal data fusion and reinforcement learning, are anticipated to elevate the capabilities of disease classification systems further and ultimately enable health professionals to receive timely and reliable diagnostic support (Umirzakova et al., 2024).

The tracking of deep learning in disease classification began with its predecessor, Artificial Neural Networks (ANN), which had already been created using machine learning techniques. Around the beginning of the 2000s, deep learning architectures with multiple hidden layers and other complexities were discovered (Sirisha et al., 2023). This method was more conducive to generating complex datasets, thereby enabling discoveries across various areas, including healthcare.

Despite this imitation of perception and layering of intelligence, our machines remain remarkably plastic, incapable of creative thinking or grasping the truth as they navigate the world's complexities. From one perspective, the future of artificial intelligence is very bright. The potential for greater insight and efficiency through data analytics, robotic automation, autonomous cars, and similar solutions is very positive (Abbaoui et al., 2024).

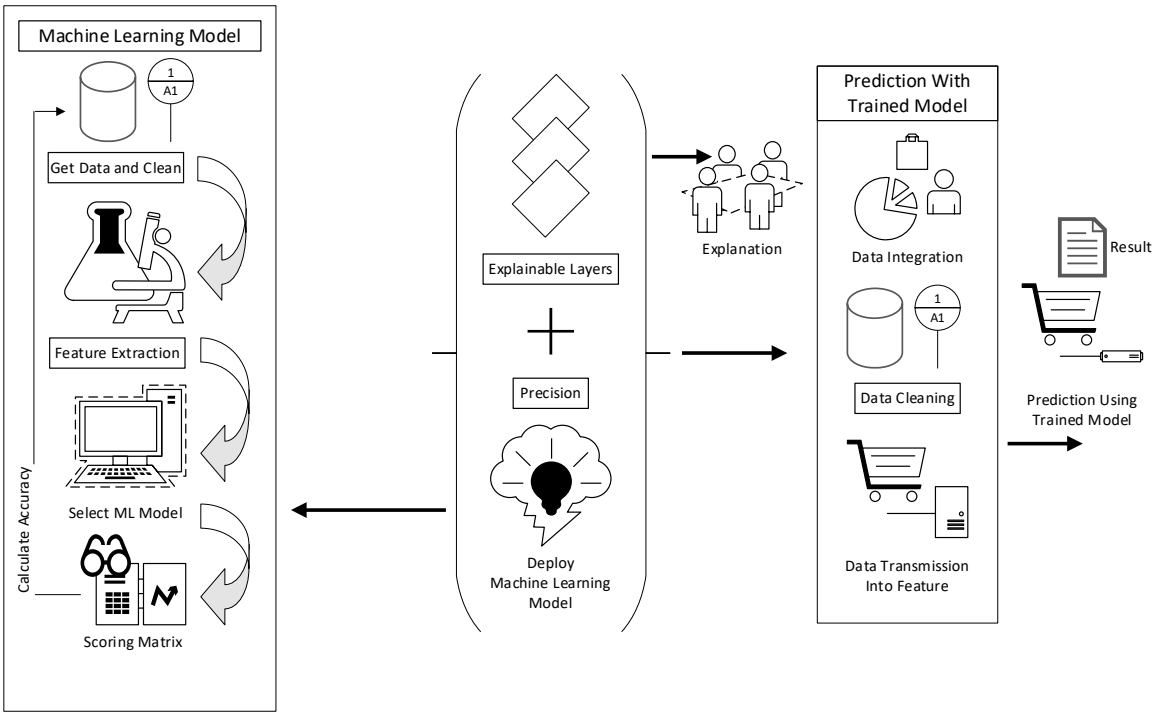
Among deep learning models, CNNs can target tumors in radiology images. Expertise in time-series analysis of patient records is a valuable tool. Current advancements have not abandoned techniques such as transfer learning, where pre-trained models are further trained on data from medical domains, and attention mechanisms that enhance model interpretation

(Lastrucci et al., 2024). GANs, or Generative Adversarial Networks, are also emerging as tools for data augmentation and anomaly detection. These accomplishments demonstrate progress, but challenges remain, including data scarcity, model interpretability, and ethical concerns. Research to date on these issues is extensive and varied, with federated learning and explainable AI being among the primary techniques employed.

Deep learning's developments in rise classification also highlight its transformative impact on the clinical decision-making process and final healthcare outcomes. Ongoing crosstalk of various disciplines and scrupulous ethics will be crucial in turning this research into accurate medical diagnostics and treatments.

Figure 5

A Conceptual Framework Machine Learning Model with Explainability



Note. The figure is driven by “An XAI-based Autism Detection: The Context Behind the Detection” by Biswas et al. (2021).

2.11 Analysis of existing studies (Related work).

2.11.1 NLP Applications in Clinical Decision Support

Recent advancements in natural language processing have transformed the analysis of clinical text data. Golinelli et al. (2020) systematically reviewed NLP implementations across diagnostic contexts and identified significant variability in performance metrics. Their analysis revealed that while NLP systems demonstrate promising accuracy in controlled settings (82-91%), performance degrades substantially when applied to diverse clinical environments with heterogeneous documentation practices. This implementation gap represents a critical challenge for translating technological capabilities into clinical impact.

Several studies have investigated specific NLP approaches for analyzing clinical texts. Wang et al. (2022) examined BERT-based models for processing clinical narratives, achieving 87.3% accuracy in symptom extraction from electronic health records. However, their analysis acknowledged significant limitations in handling domain-specific terminology and the frequent negation patterns in clinical documentation. Similarly, Chen and Rodriguez (2021) evaluated transformer-based architectures for processing unstructured clinical notes, demonstrating superior performance over traditional machine learning approaches but identifying persistent challenges in contextual interpretation of temporal relationships between symptoms.

2.11.2 Audio Analysis in Diagnostic Applications

Complementary to text-based approaches, audio analysis systems have become valuable tools for capturing diagnostic indicators. Kumar et al. (2023) developed a CNN-based classifier for respiratory sound analysis, achieving 83.5% accuracy in distinguishing among five respiratory conditions. Their comparative analysis revealed that spectral features outperformed temporal features, suggesting specific architectural considerations for clinical audio processing. However, they identified significant limitations in real-world

implementation, particularly regarding ambient noise interference and variability in recording quality across clinical settings.

Research by Martinez and Ali (2021) on speech pattern analysis for neurological condition assessment demonstrated promising results (79.1% sensitivity and 82.4% specificity) but highlighted critical challenges in creating generalizable models. Their work emphasized the importance of diverse training data encompassing demographic variability, a limitation acknowledged across multiple studies in this domain. This finding highlights the importance of methodologically rigorous approaches to dataset construction that accurately reflect clinical diversity.

2.11.3 Integrated Multi-modal Approaches

Emerging research suggests that integrated approaches combining multiple data modalities may overcome limitations inherent to single-modality systems. Zhang et al. (2023) developed a hybrid classifier integrating text and audio features for psychiatric evaluation, achieving a 12.7% improvement in diagnostic accuracy over the best-performing single-modality system. Their analysis indicated that different modalities captured complementary aspects of symptom manifestation, with textual data excelling in semantic content analysis while audio features better captured emotional and behavioral indicators.

Despite these promising results, Ramirez and Singh (2022) identified significant methodological challenges in multimodal integration, particularly in temporal alignment across data streams and in developing appropriate fusion strategies. Their comparative analysis of early, late, and hybrid fusion approaches revealed that architectural decisions have a significant impact on performance across different diagnostic categories, suggesting that optimal integration strategies may be condition-specific rather than universally applicable.

2.11.4 Research Gaps and Opportunities

Critical analysis of existing literature reveals several significant gaps that this study addresses:

1. **Methodological limitations in evaluation**: Current research predominantly evaluates technical performance in idealized settings, with insufficient consideration of real-world clinical conditions. This study implements a comprehensive evaluation framework that assesses technical performance and practical implementation factors.
2. **Insufficient attention to generalizability**: Many studies demonstrate proof of concept with limited datasets that inadequately represent clinical diversity. This research employs rigorous cross-validation approaches using demographically diverse data to address concerns about generalizability.
3. **Limited integration of implementation science**: Technical capabilities frequently overshadow adoption considerations in the existing literature. This study explicitly addresses factors influencing clinical implementation beyond technical performance by incorporating Rogers' Diffusion of Innovation framework.
4. **Architectural optimization for clinical contexts**: Previous works often apply general-purpose architectures without sufficient adaptation to clinical requirements. This research develops specialized architectural components optimized for the unique characteristics of medical data.

By addressing these gaps, this study extends beyond incremental improvements in classification accuracy to establish a comprehensive framework for developing and evaluating clinically viable computational diagnostic support systems.

2.11.5 Data Monitoring for Health Data Using Internet of Medical Things (IoMT) and Random Forest Classifier.

Internet of Things (IoT) sensors will collect data from the patient and the surrounding environment. The data must be properly sanitized to improve the performance of the random forest classifier. The classifier will ensure that all smart devices are properly configured to collect accurate data. The proper utilization of data can have a significant impact on the world's healthcare system. It makes it feasible for monitoring to reach those who need easy access to an efficient health observation system.

The data obtained is then analyzed using different classifiers, which helps predict the classifier's accuracy. Shah (2020) noted that once the output is deployed, the classifier's predictions will be relayed wirelessly to medical specialists, who can offer valuable recommendations. These circumstances are also available, and we aim to improve them by using historical data to predict future problems through prescriptive analytics. This will enable doctors to use appropriate surgeries based on the classifier's predictions.

The study review identified that various machine learning methods, using publicly available datasets on a cloud platform, were employed to build a real-time, remote health monitoring system that leverages IoT infrastructure and cloud computing. The research recommendation was based on textual and empirical data available in the cloud. The study proposes a model that reveals a mechanism by which information from the public database can help identify patterns in health data (symptoms) and inform trustworthy decisions (Khatoon, 2020). The study has developed another classifier to compare with the existing one, but it has been identified as having the best performance.

The classifier can quickly predict diseases such as heart disease, breast cancer, and diabetes by using a variety of input attributes specific to each disease. Another classifier,

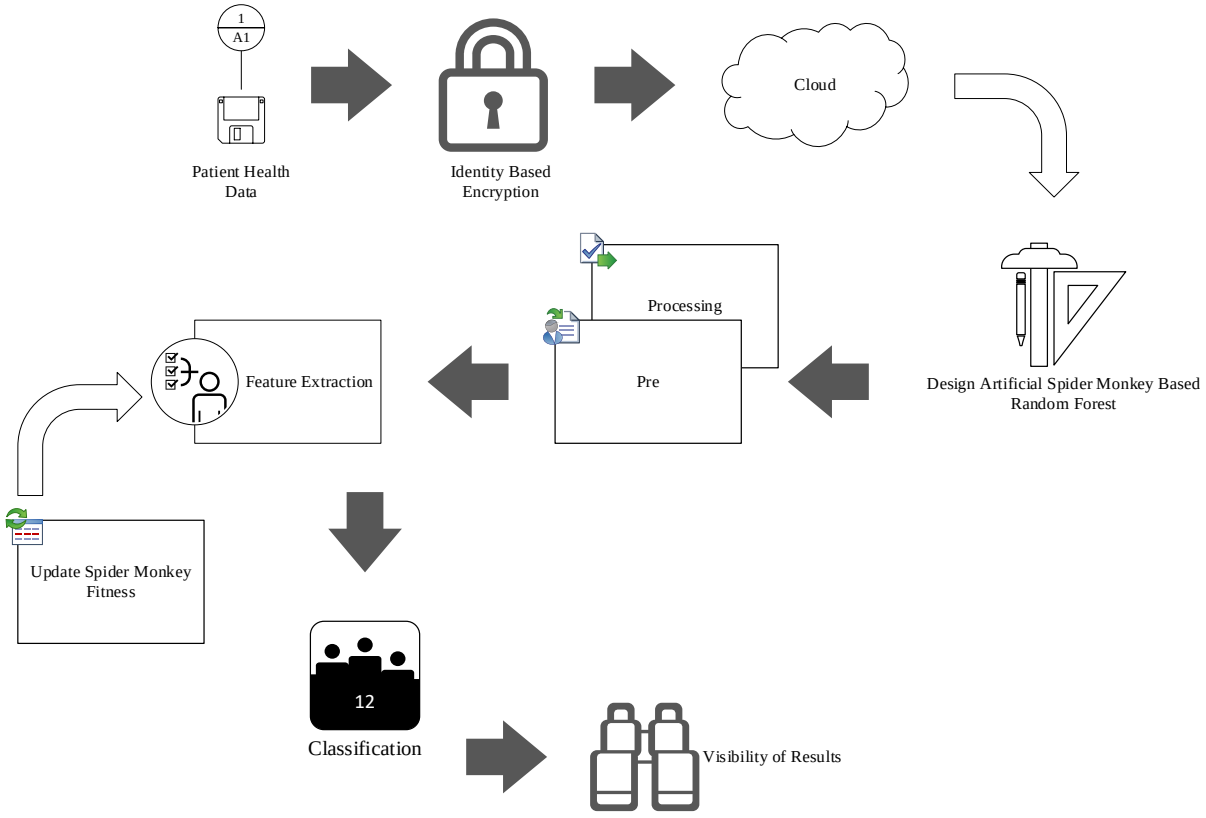
designed to aid prediction, was KNN, which achieved an accuracy of 61%. Support vector machine (SVM) achieved an accuracy of 61.38%. The decision tree algorithm achieved an accuracy of 68.08%. The classifier from the study correctly predicted the data used to train the model. The trained model achieved high performance due to the following aspects exhibited by the classifier (Dey et al., 2020). The classifier was less prone to being influenced by outliers. The output differed when applied to another classifier, such as an SVM or a decision tree classifier. The classifier's ensemble nature helps mitigate the impact of individual data-point skewness.

They can effectively handle overfitting. The ensemble nature of Random Forests helps to mitigate overfitting, a common issue with individual decision trees. The algorithm constructs multiple trees and combines their predictions, yielding a more robust, generalized model. They exhibit high performance. They perform well and provide high accuracy during the task classification process. They reduce the risk of overfitting and capture complex relationships in the data.

A developed framework based on Artificial Spider Monkey-based Random Forest (ASM-RF) is a hybrid that leverages the predictive analytics of Random Forests and artificial intelligence to inform decisions based on patient health data. The hybrid framework uses a fitness function to evaluate the spider monkey's performance and calculate its accuracy. The classifier can effectively support proper decision-making in the healthcare system. The classifier achieved 99.52% accuracy and 99.12% precision, and the model ran in approximately 6 seconds.

Figure 6

A representation of the architecture of the proposed Artificial Spider Monkey-based Random Forest Hybrid Framework highlights the interlink between the different components in the healthcare system.



Note. The figure used in “A Novel Artificial Spider Monkey-Based Random Forest Hybrid Framework for Monitoring and Predictive Diagnoses of Patients' Healthcare” by Shahid et al., 2023.

Meanwhile, the AlexNet model, which focused on treating all diseases, applied a hybrid framework and achieved high accuracy. REYAZU is an IoMT cardiac-based management system that enables remote monitoring of a patient's wellness without requiring human intervention.

2.11.5.1 Weakness Exhibited by the Classifier in the Study

This classifier was found to be sensitive to the data collected, as the data significantly impacted its performance. The classifier's performance was lower than that of other machine

learning classifiers. This ensures rigorous data preprocessing and exploratory analysis to mitigate critical issues in feature selection.

Feature importance dilution, where the medical records were highly correlated. Data obtained from the research sensor were found to be significantly correlated, with similar symptoms observed in patients with similar diseases (Martínez Munoz et al., 2022). Therefore, the classifier shows that when many symptoms were interrelated, it was difficult to interpret during visualization, making the feature's importance less meaningful.

The random forest classifier was inappropriate for streaming data because it required multiple decision trees to produce a single result, making it challenging to update the learned function with new data. Rebuilding a classifier is extremely expensive if new data makes it less accurate. This is a gap for this model, as new diseases continue to emerge and mutate due to dynamic changes in various DNA sequences.

Therefore, ensuring that the classifier remains practical and continually enhanced is essential. The model we develop must address these weaknesses and ensure that the new features are effectively incorporated into the knowledge base.

2.11.6 A Survey of Audio Classification Using Deep Learning

Zaman et al. (2023) developed deep-learning models for various audio classification tasks. In this instance, the corpus is addressed as a five-deep neural network architecture. In particular, works with CNNs, RNNs, Autoencoders, Transformers, and Hybrid Models (hybrid deep learning models and hybrid deep learning models with traditional classifiers) are central to the study (Nanni et al., 2021). The CNN model can differentiate sound signals into speech, music, and environmental sounds. They may also serve non-lingual purposes such as speech recognition, speaker identification, and emotion recognition.

RNNs have myriad applications, such as audio classification and segmentation. RNNs can work with streaming data and thus may be trained to capture trends of audio stream signals. Such models can later label audio streams into different classes (Wei et al., 2020). Therefore, an autoencoder can be used to study the features of audio signals and classify them into different sounds. The Transformers are perfect for audio classification, as deep learning algorithms can be implemented to separate the encoder and decoder.

These audio signal features can be extracted and identified using time-domain and frequency-domain analysis. The audio classification models combine diverse deep learning architectures (CNN-RNN) with standard machine learning techniques (Tsalera et al., 2021). Combining various deep learning models allows the advantages of one to negate the inadequacies of others. This discussion surveyed the literature in detail across various categories.

2.11.6.1 Weakness of the Study

The assessment of Audio Classification via Deep Learning failed to incorporate the latest developments, favoring outdated approaches. As for the tested sample sizes, this can be an additional factor contributing to problems with transferability (Koutini et al., 2021). Moreover, it will mostly fail to account for dilemmas such as data shortage and unique circumstances that compromise the effectiveness of such models. Additionally, comparing deep learning methods to other methods can be crucial for gaining a broader understanding of their performance and limitations (Nanni et al., 2020). Overcoming these limitations by thoroughly updating the literature review, detailing the methodology, and providing broader contextualization will enhance the depth and reliability of the research.

2.11.6.2 Strength of the Study

The deep learning modalities covered in this audio survey system include all classification methods used for audio classification. By combining various methods, strategies, and experiments, the assessment of significant technological changes in this area has become more accurate (Nanni et al., 2021). Because the extensive representation of worked-out measures, including speed and accuracy, is involved, the reliability of the research becomes greater. In the 'difficulty and future direction' part of the statement, we will realize that the statement is intended to introduce new concepts to researchers and practitioners, which is helpful (Zhang et al., 2024). To sum up, it is easy to conclude whether this method matches the box.

2.11.7 Audio Classification Using Braided Convolutional Neural Networks

Audio classification using braided convoluted neural networks primarily focuses on adopting convolutional neural networks to advance existing techniques (Mahmood et al., 2023). The relatively unprecedented success gave rise to the extension of CNNs to the domain of sound data. Although recent publications position hidden Markov models and deep neural networks to extrapolate audio information, there is still much to explore and discover. This study aims to do audio classification based on the given frequency of the chart and then use a CNN-based architecture for classification (Sinha et al., 2020).

The paradigmatic neural network architecture is a convolutional neural network (CNN), designed to mimic the sparsity characteristics of mammalian auditory cortex receptors (Ali et al., 2022). The viability of the proposed material is evaluated on regular benchmark datasets for audio classification. For instance, the Google speech commands datasets (GSC v1 and GSC v2) and the UrbanSound8K dataset (US8K) (Anh et al., 2021). The proposed CNN architecture, which includes a braided convolutional neural network, achieved average recognition

accuracies of 49.15%, 95%, and 91.9% on the GSCv1, GSCv2, and US8 K datasets, respectively.

2.11.7.1 Strength of the Study.

Another contribution is that the authors introduce a multiple convolutional neural network (CNN) for audio-video data classification. Thus, deeper multi-channel sound processing will be enabled, making music transmission more dynamic and vibrant. The braided CNN pattern is well optimized to detect feature dissimilarity, whether local or global, thereby improving the classifier's accuracy. That is AI's key strength: its ability to learn and operate continuously across all audio environments. The approach is considered more fundamental than sophisticated and is more successful in terms of the uniqueness of the classifier's efficiency compared to traditional CNN architectures. Such experiments introduce new products in audio management that leverage high-performance Brixed CNNs, thereby stimulating the design of new applications, such as speech recognition, music analysis, and medical diagnostics.

The classifier is robust to outliers: the three-decision tree classifier can efficiently handle them. This could be attributed to the data not being certified by the classifier, but the three-level classifiers can handle them efficiently (Jeong et al., 2022). An error may have occurred during data collection or in unfamiliar patient conditions. Decision tree classifiers are less sensitive to outliers than other machine learning models when benchmarked.

The development of machine learning-based diabetes prediction and intelligent web applications was conducted by Nazin et al. (2020). The study involves several feature selection techniques and the identification and ranking of risk factors. The performance of two different data sets was evaluated in exhaustive experiments. In contrast to some recent studies, the model was compared for accuracy using different machine learning (ML) algorithms across various datasets. The results show that the proposed model achieves higher accuracy, ranging from

2.71% to 13.13%, depending on the dataset and machine learning algorithm used. Lastly, a machine learning algorithm with the highest accuracy for the research is chosen. The association model formulation serves as the holotropic facet of the approach, integrated into a web application utilizing the Python Flask web development framework.

2.11.7.2 Weakness of the Classifier

This type of classifier demonstrated that the output cannot be guaranteed when the data are imbalanced. This shows that decision trees can be biased, especially when the class becomes dominant. This is because they tend to prioritize the majority class, leading to a performance that is not optimal, especially for minority classes.

The association between tuberculosis and COVID-19 could not be established using the decision tree with three classifiers. This happens especially if those relationships are not easily represented by axis-aligned splits (Emmanuel et al., 2021). Models such as random forest or gradient boosting can quickly identify associations between the classifiers.

Decision trees employ a greedy, recursive approach to local optimization, which can sometimes result in a suboptimal overall tree structure. The algorithm selects the optimal split for each node at the time without considering the potential consequences for future nodes.

The classifier is highly sensitive to outliers, noisy data, and errors during model training. Small changes in the data can lead to a dramatically different decision about the tree structure. It is not as expressive as other machine learning models. It exhibits challenges where features and target variables are complex, and this is exacerbated by non-linear data, making this classifier highly versatile.

Experiments on audio samples different from those provided would be required to remove the doubt of whether the work of the paper "Audio Classification Using Braided Convolutional Neural Networks will face generalizability limitations. In audio classification

tasks, CNNs are used to extract acoustic features, yielding results that depend on feature and hyperparameter settings and reflect the dataset's size.

AI practitioners may face challenges with data interpretation and excessive resource consumption when using differentiable reset mechanisms with advanced braided CNN structures. Moreover, the final trial may not appear sufficiently intense compared with other advanced methods of braided CNNs, which makes understanding braided CNNs more complex in other possible audio classification contexts beyond the current one.

2.11.8 Liver Disease Prediction Using SVM and Naïve Bayes Algorithms

Researchers have consistently faced challenges in the healthcare sector when attempting to predict the onset of certain fatal diseases from large medical record databases. Nowadays, technology is essential in the healthcare sector. Accurate visualization can be derived from the data, and effective decision-based medicine can result. Therefore, data mining techniques can be fully applied to medical data to enhance predictions through appropriate classification, clustering, and association rule analysis, thereby helping identify patterns in the data. These patterns help in understanding the data from a standpoint.

Data mining techniques have been employed in various classification models to provide insights into the data and facilitate diagnosis. The study proposes a hybrid classifier combining support vector machines and Naïve Bayes to predict liver disease (Albahri et al., 2020). Therefore, the research's primary objective is to predict multiple diseases, including Cirrhosis, Bile Duct, Chronic Hepatitis, Liver Cancer, and Acute Hepatitis, from the Liver Function Test (LFT) dataset using the aforementioned hybrid classifier.

Sensitive organs are vital because they perform serious bodily functions and require serious care to avoid failure. The overall functionality of these organs is unique, as they help coordinate the bodily functions of other organs.

Aspects such as sugar regulation and red blood cell breakdown are metabolic processes, essential activities in the body that require careful management to operate optimally. The liver, weighing approximately 3 pounds, performs many crucial functions related to digestion, metabolism, immunity, and nutrient storage. This vital organ plays a crucial role in maintaining tissue health, as the body would quickly suffer from a lack of energy and nutrients without proper functioning. A variety of factors contribute to the potential development of liver disease.

Support vector machines have existed since 1975 and have helped classify components that would have taken centuries to classify (Georges & Seckin, 2022). It has facilitated the implementation of various projects, helping solve real-life situations. The classifier is efficient in both regression analysis and classification.

Implementing a non-linear mapping on the original training data transforms it into a higher dimension, allowing for the search for an optimal separating hyperplane. This hyperplane effortlessly distinguishes between the data from two distinct classes. Furthermore, by including support vectors and margins, the SVM can find and fully exploit this hyperplane. With its ability to maximize the margin and accurately classify both classes while minimizing errors, the SVM is versatile and excels at data classification (Yousefi et al., 2021). While it can also be applied to other optimization problems, such as regression, the SVM's primary task is to classify data effectively. The data points are labeled as positive or negative, and the goal is to find a hyperplane that maximally separates them. Here, one classifier fails to perform optimally, resulting in incorrect classification.

The research was focused heavily on the classifier's execution time and performance. SVM was well-suited for this study. SVMs also demonstrated improved classification times for liver disease and other conditions. Another classifier that was compared in this research exhibited the following accuracies. The accuracy measure for five classifiers includes.

Those classifiers, however, have produced better results in classifying medical ailment symptoms, which ensures that the application is well thought out. Once deployed, it can help predict components. Ideally, the five classifiers achieve high performance across the evaluation metrics.

The research can be improved by adopting a high-accuracy classifier with appropriate data, which requires proper validation and sanitization. The level of true negatives from the classifiers used was 38.73 for the Multinomial Naive Bayes, 20.34 for the Support Vector Machine, and 38.84 for the Feedforward Neural Network (FNN).

Therefore, the research reveals numerous gaps that could be addressed to improve the classifiers' performance and facilitate accurate predictions from the data. Exploratory data analysis is necessary to ensure data sanitization (Pandey et al., 2023). Removing data inconsistencies, such as null values, duplicate values, incorrect labeling, and missing values, improves the model's accuracy. In a real scenario, training the two classifiers is computationally expensive. Therefore, a single classifier that achieves higher performance must be trained, and quality-sanitized and normalized data must be used.

Sazzadu et al. (2023) conducted a comparative study on liver disease predictions using all supervised machine learning algorithms. The research aims to identify the most suitable classifier to help mitigate the effects of choriocarcinoma liver infections on patients. The research employs six supervised machine learning classifiers. The classifiers' accuracy was unexpected, as some could have achieved an average of 75% (SVM). At least one achieved 53% accuracy (decision tree), demonstrating that, with proper effort to improve training data quality and address issues related to classifier variability in training datasets, it is well validated. Consequently, it will lead to better performance and become very helpful, especially in an environment where we have limited liver specialists.

Panib and Kumar developed a rule-based data-mining algorithm to assist the healthcare sector in making informed decisions about liver disorders. The classifier could discover hidden patterns in the patient's medical history and make appropriate decisions. Natural language processing work at this point, in which thematic and sentiment analysis were performed to ensure all activities were achieved optimally, would help the medical institution make informed decisions based on the support system to assist doctors. The classifier used was a decision tree, which achieved an accuracy of 78%.

2.11.8.1 Strength of the Classifier

Using non-linear kernels in SVM models can make decision-making difficult to comprehend. In the healthcare industry, where clinician trust is essential and aligning model decisions with medical knowledge is critical, interpretability is crucial. Furthermore, obtaining reliable results from imbalanced health record datasets with significant class imbalance can be challenging for SVMs (Silva et al., 2020). This may lead to prioritizing the majority class, thereby hindering the performance of the minority class.

Streaming data is not ideal for SVMs, as they are trained only on a fixed dataset; therefore, incorporating new data in real time becomes challenging for the classifier. This can pose a problem when health records are constantly updated, and the model must evolve accordingly.

SVMs require specific hyperparameters that must be finely tuned for optimal results. There are automated approaches to identify the best hyperparameters, as their selection significantly affects SVM performance.

2.11.8.2 Weakness of the Classifier

SVMs have earned a reputation for delivering impressive results, but their models can be challenging to interpret, especially when advanced kernel functions are employed. Grasping

the decision boundaries within a high-dimensional space can prove daunting. Tweaking SVMs for optimal performance requires fine-tuning regularization and kernel parameters.

However, SVMs' sensitivity to these choices means that finding optimal values can involve significant trial-and-error. In dealing with enormous datasets, SVMs' computational demands can be prohibitive (Jalilov et al., 2021). In such cases, it may be more prudent to turn to alternative algorithms, such as stochastic gradient descent, which can better handle large datasets.

2.11.9 Deep Learning-Based Decision-Tree Classifier for COVID-19 Diagnosis from Chest X-ray Imaging

The availability of medical data has enabled various healthcare sectors to achieve more significant milestones in the era of technology and artificial intelligence. Therefore, processing the data will ensure that appropriate data has been collected and help deduce essential insights. However, there are significant ways to use the data to train a classifier that can help treat a severe disease.

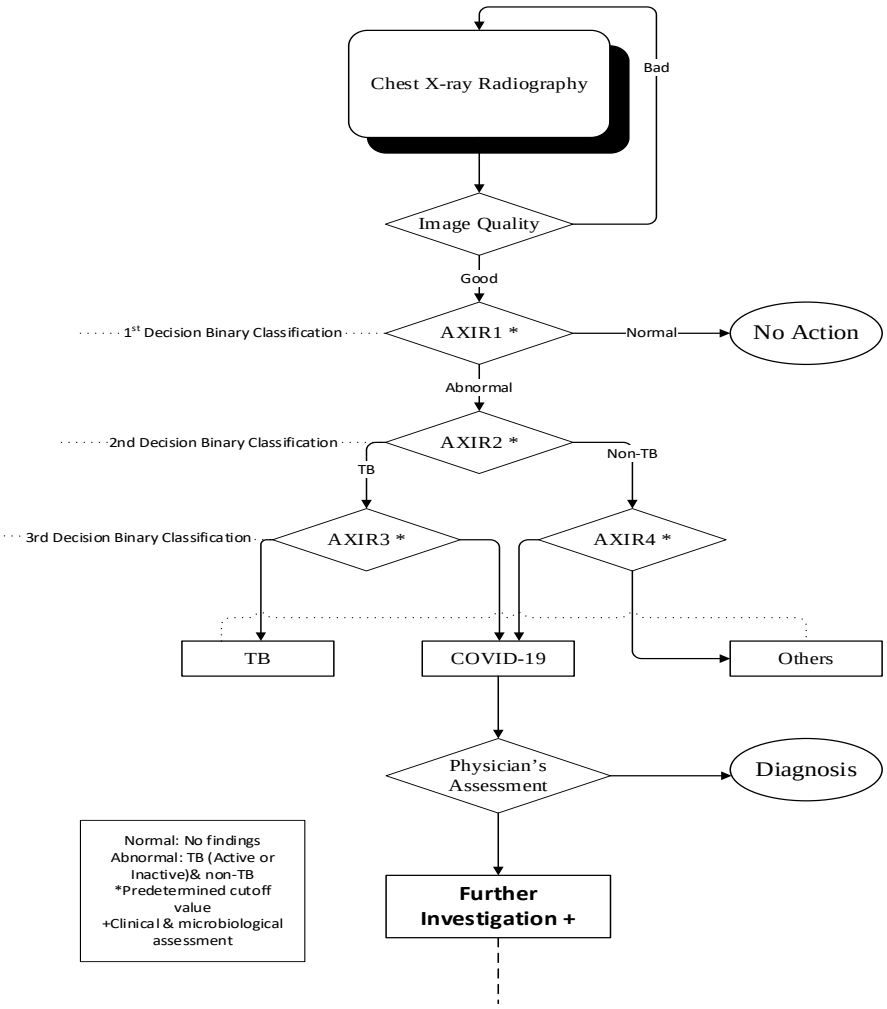
This study has combined data mining and traditional data analysis approaches to predict the causes of COVID-19. This highlights how data mining in the medical field is a significant area that requires efficient configuration to enable accurate decisions from the data.

Effectively categorizing and forecasting medical data presents various obstacles in Medical Data Mining. According to global statistics, COVID-19 is the primary cause of mortality (Yoo et al., 2020). Predicting the spread of COVID-19 is challenging for medical professionals, requiring extensive expertise and technical knowledge. However, within the vast expanse of the healthcare industry, invaluable insights lurk that can inform critical decision-making. By utilizing data mining techniques such as decision trees and Naïve Bayes, this study successfully predicts COVID-19 disease with an outstanding accuracy of 95%. Data mining

empowers the healthcare sector to identify patterns in large datasets, enabling targeted predictions.

Figure 7

Workflow determines whether the chest X-ray image shows an average, tubercular (TB), or COVID-19-infected lung. AXIR (Automated X-ray Imaging Radiography system).



Note. The figure used in "Deep Learning-Based Decision-Tree Classifier for COVID-19 Diagnosis From Chest X-ray Imaging" by Yoo et al ., 2020.

The proposed classifier in this study utilizes a deep learning-based decision tree classifier to predict COVID-19. COVID-19 was one of the fatal diseases that killed large groups of people in the years 2020, 2021, and early 2022 (Allioui & Mourdi, 2023). Therefore, concerns have started to arise about how to treat the deadly disease to reduce its highly contagious nature and the level of death.

This proposed model will contain three binary decision trees. Deep learning with a CNN implemented in the PyTorch framework will explicitly train all models. The three

classifiers had to make predictions based on the available data, including the first decision tree and torsion tree, and classify CXR images as normal or abnormal. The second decision tree classifier aims to identify whether the abnormal images contain elements of tuberculosis. The third decision tree classifier will help determine whether the classifier contains any COVID-19 features. The accuracies of the classifiers were 98% and 80%, respectively, while the third classifier had an accuracy of 95%. The proposed classifier was a deep learning-based decision-tree classifier that could be used in pre-screening patients to facilitate triage and expedite decision-making before RT-PCR results are available.

The data used in the research had not been certified; hence, it required further consideration. There was considerable bias because the data represented only a small sample of patients with COVID-19—1,750—yet hundreds of thousands of patients have been tested for this disease.

The validity and reliability of the data could not be ascertained because the training data could not be confirmed against the pathological data. Therefore, the data may not accurately reflect the actual situation during the COVID-19 pandemic (Yoo et al., 2020). Without pathologically confirmed data, the model's results are unreliable. Therefore, predicting new cases requires new pathological data.

The obvious challenge in implementing the classifier is that three-level decision classifiers will be computationally expensive, taking three times as long as a single classifier. All these could have been achieved using three classifiers, thereby reducing classification time. Furthermore, a wide range of data augmentation techniques exists for image data. To enhance the effectiveness of our deep learning model, we augmented our training data using simple yet effective techniques, such as horizontal flipping, rotation, and shifting.

More advanced techniques, such as stochastic region sharpening, elastic transforms, and randomly erasing patches, could be explored to improve the model's performance further. It is crucial to continue researching advanced augmentation techniques to develop more sophisticated models and create a system capable of producing reliable statistical models with limited training data (Ennab & Mcheick, 2022).

From the study, it can be seen that the classifier focused primarily on the performance of the three classifiers, which is beneficial because the accuracy of most classifiers determines their efficiency. According to the research, using a multi-three-layer classifier is unnecessary, as it does not yield greater accuracy in predicting the data and the classifier.

The study encompassed numerous comparisons across various diseases, rather than focusing on a single primary disease, COVID-19. Rather than using tuberculosis as the indicator for COVID-19, the classifier should have selected all the features that could constitute COVID-19. Therefore, the approach is more time-consuming, making the classifier less effective (Russo et al., 2021).

Research on artificial intelligence techniques for COVID-19 using chest X-rays with convoluted neural networks. According to Saddam et al. (2020), the classifier's accuracy determines its efficiency and performance. The research achieved an accuracy rate of 96%. Unlike the previous classifier, this research heavily relies on large datasets to achieve better performance and accuracy. The classifier's strength was that it fully fine-tuned the parameters and applied them to the classifier to improve results. Ideally, large datasets help make more informed decisions.

Adopting the automated analysis method will save doctors a great deal of time. The paper proposes and implements an intelligent CNN approach. The hyperparameters of CNN are optimized by multi-objective adaptive differential evolution (MADE). Extended

experiments were conducted using the benchmark COVID-19 dataset. The comparison of approaches reveals the superiority of the suggested approach over competitive machine learning models in various criteria. The optimized CNN achieved an accuracy of 97.2%.

2.11.9.1 Strength of the Classifier

The classifier was robust to outliers: The three-decision tree classifier could efficiently handle the outliers. This could be identified as the data collected could not be certified by the classifier, but the three-level classifiers could handle them efficiently. This could have been occasioned by an error during data collection or unfamiliar patient conditions (Križanić, 2020). Decision tree classifiers are less sensitive to outliers than other machine learning models when benchmarked.

The classifier handles all types of data, including both numerical and categorical data. This process does not require extensive data preprocessing, enabling proper handling of patient demographics, lab results, and categorical variables, such as diagnoses.

Decision trees can handle missing values in the dataset without requiring imputation. In medical data, missing values are not uncommon for various reasons, and working with such data is advantageous.

The classifier could handle missing values even without requiring any imputations. However, missing values are not a common issue in this dataset, which makes working with such data very advantageous.

The three classifiers efficiently handled the voluminous data in this study, though execution time was high. Given the dataset's size, the classifier's scalability was significant, and it was therefore crucial to efficiently handle the data.

2.11.9.2 Weakness of the Classifier.

This type of classifier showed that the output cannot be guaranteed when the data are imbalanced. This shows that decision trees can be biased, especially when the class becomes dominant. This is because they tend to prioritize the majority class, leading to a performance that is not optimal, especially for minority classes.

The association between tuberculosis and COVID-19 could not be established using the decision tree with three classifiers. This happens especially if axis-aligned splits do not easily represent those relationships. Models such as random forest or gradient boosting can quickly identify associations between the classifiers.

Decision trees employ a greedy, recursive approach to local optimization, which can sometimes result in suboptimal overall tree structures (Bertsimas & Öztürk, 2023). The algorithm selects the optimal split for each node at the moment without considering the potential consequences for future nodes.

The classifier is highly sensitive to outliers, noisy data, and errors during model training. Small changes in the data can lead to a dramatically different decision about the tree structure. They are not as expressive as other machine learning models. They exhibit challenges when features and target variables are complex, particularly when the data is non-linear, making this classifier highly versatile.

2.11.10 Classifying Alzheimer's Disease Using Audio and Text-Based Representations of Speech.

In this study, a gender- and age-balanced ADReSS dataset will be utilized to develop an automated classification model for Alzheimer's disease (AD) assessment from spontaneous speech recordings. Their performance was compared to a neural network implemented in a convolutional layer and another in long-term memory(Haulcy & Glass, 2021), and this was

done after having trained the networks on audio features (i-vectors and x-vectors) and text features (word vectors, BERT embeddings, LIWC features, and CLAN features).

Audio and text features were used to train five regression models, as researchers suggested, explicitly designed to predict each patient's Mini-Mental State score (0-30). The highest result in the experiment was achieved by a set of the best classifiers, specifically the support vector machine and random forest classifiers with BERT embeddings, which attained an accuracy of 85.3% on the test set. The highest-performing regression model on the test dataset was a gradient boosting regressor trained on BERT embeddings. CLAN features obtained a root mean squared error through the execution of both conduct; effects of speaking have been shown for AD classification and estimation of neuropsychological scores.

2.11.10.1 The Strength of the Study

The study's approach utilizes a classifier that operates on both text and audio. Finding other research methods that are more panoramic in their presentation, such as presenting 2D views from different angles of the structural models of Alzheimer's disease, secures a holistic view of the matter under investigation. Integrated research represents a major step toward developing new, more accurate diagnostic tools and detecting diseases early. When such a mechanism is in place, early disease detection is possible, and treatment is therefore not only available but also rapid. The study aims to bridge fields such as speech processing, machine learning, and personal healthcare, emphasizing multidimensional solutions to provide a comprehensive picture of healthcare today.

2.11.10.2 The Weakness of the Study

The study identified Alzheimer's disease using speech-based approaches, which could be attributed to the classifier used and the lack of proper data preprocessing for the data used in the application's prediction. Therefore, this gap should be implemented effectively to ensure

the classifier produces proper results. Limitations highlight the importance of stricter validation, more extensive and diverse datasets, and a careful examination of the results of these types of studies.

2.12 Alternative Frameworks, With a Justification of Why the Selected Framework Was Chosen.

Another framework could have been applied in this research, but a conceptual framework was selected. Other frameworks include, but are not limited to, theoretical frameworks and hybrid theories that combine conceptual and theoretical models. (Alowais et al., 2023) “AI can diagnose diseases, develop personalized treatment plans, and assist clinicians with decision-making. Rather than simply automating tasks, AI is about developing technologies that can enhance patient care across healthcare settings.” However, the conceptual model has been chosen because it helps document the literature using factual information from the reviewed articles rather than theories. They critique the research work, enhancing a proper understanding of domain knowledge, especially in natural language processing and deep learning analysis.

2.13 Describe How and Why the Selected Framework Relates to The Present Study

The conceptual framework was chosen because it enables the establishment of cause-and-effect relationships for any phenomenon in a review, thereby helping the researcher develop the capacity to understand the work effectively and providing a conceptual foundation for the research. The framework chosen depends on the thesis's structure and the primary information to be analyzed.

The researcher can synthesize and organize existing knowledge within the problem domain. Adopting a conceptual framework in this research will enable the researcher to develop a classification approach for text and audio data to diagnose and treat diseases, drawing on NLP and D.Phils. Categorization and organization enable concepts to be organized into theories, themes, and methodologies, thereby enhancing the structured review of articles across various study domains.

Once proper themes have been identified, the researcher can develop research questions, research gaps (or problem statements), and purpose statements, as the reviewed articles will reveal gaps, inconsistencies, and contradictions in existing studies. This theoretical analysis and organization approach offers an opportunity for critique, which is essential for advanced scholarly discourse and contributes effectively to the body of knowledge in the field under study.

2.14 Summary

This section identifies critical themes in the reviewed articles, highlights gaps, and supports the study. The articles reveal challenges that have plagued the health sector and require proper implementation to enable advanced disease diagnosis that doctors could not previously identify. Inaccurate diagnoses of deadly diseases can wipe out the entire community. The classifier's performance can be observed in the various studies identified. A proper classifier produces proper output, meaning the classifier's intention is identified.

Identifying critical areas of concern across all reviewed articles revealed poor model training, resulting in lower accuracy; poor dataset labeling; and extrapolation of data with null values, leading to inaccurate diagnoses, especially for X-rays used in treatments. Additionally, insufficient sample datasets were used for model training.

In reviewing all the studies, the issue of proper data to feed the model is a major concern. The research yields several divergent opinions, and once these views are implemented, the model output will be enhanced. The choice of the classifier has been identified as having a significant impact on its accuracy. This has been discussed in depth in the classifier provided.

The data used to train the classifier has been a critical factor. High-quality datasets yield better performance for any classifier. All the reviewed papers have used classifiers to predict the data, and to a greater extent, some classifiers have used a single algorithm to train the model. However, we have identified two classifiers for computationally intensive and time-consuming tasks. They demonstrated that an appropriate classifier was applied and achieved accurate performance, but the medical data used was only textual. However, the patient's medical history contradicts the information available in the system. Therefore, audio data from the patient's narrative can be analyzed. Therefore, this is a gap that needs to be addressed, as the audio data have not yet been discussed in all the studies reviewed.

In analyzing the mentioned aspects and identifying gaps, an appropriate research design and methodology should be applied to ensure that all challenges in the context are addressed. Therefore, the study employs a constructive research design and machine learning (ML), including deep learning and natural language processing (NLP), for sentiment analysis to enhance research audio and text analysis from patients' medical histories or narratives. This research design and methodology align with these frameworks, as they emphasize the potential spread of NLP and DL in the healthcare industry, driven by the novelty of these technologies (Bianchini et al., 2020). These unified models also guide research on the spread of NLP and DL innovations within the complex healthcare system.

Chapter 3: Research Methodology

This chapter outlines how the project will be implemented in the following and subsequent chapters. This study addresses a critical gap in clinical decision support: the limited availability of validated computational tools for analyzing multimodal patient data. Specifically, Lu et al. (2020) identified significant challenges in medical diagnosis and treatment stemming from inadequate systems for reliably interpreting textual narratives and auditory information in healthcare settings. This deficiency impedes accurate symptom classification and contributes to diagnostic uncertainty in clinical practice. In modern healthcare, where data is abundant, failing to utilize it leads to delays, errors, and missed opportunities for early intervention. Current diagnosis and treatment methods rely on human interpretation, making the management of medical records and audio recordings complex and unpredictable. Healthcare personnel, patients, organizations, and society are affected by this issue. Data capture and storage systems are predominantly retrospective and paper-based, making it inefficient to retrieve and use the data to inform care decisions (Heys et al., 2022).

The problem of exhaustion and a decline in care quality is evidenced by the difficulties healthcare providers face in processing massive amounts of textual and auditory data (Stark et al., 2018). Consequently, patients face challenges obtaining prompt and accurate diagnoses, resulting in subpar treatment outcomes. The rising costs and potential legal risks healthcare organizations face are significant factors in the escalation of healthcare costs and the deterioration of social well-being.

This research primarily aims to develop deep learning classifiers and natural language processing models to support patient diagnoses and treatment using text and audio data. This project aims to address the inefficiencies outlined in the problem statement by utilizing advanced technologies to streamline healthcare diagnostic and treatment processes. The

problem outlined — namely, the underutilization of textual and audio data in healthcare, which leads to delays in diagnosis and incorrect treatment — is directly addressed in this study by developing models that leverage text and audio to classify patient symptoms.

The issue impacts patients, healthcare workers, and society. Inefficient medical data analysis leads to patient suffering, treatment delays, increased healthcare expenses, and legal issues. The total stakeholder influence and complexity of variables preventing data use must be clarified. Neglecting the issue risks patient suffering, increased costs, and missed opportunities for early intervention. The study will highlight the importance of answering these questions to enhance healthcare by decreasing unintended consequences and improving medical data analysis.

NLP and DL approaches in medical diagnosis and therapy will be examined using performance metrics such as F1-score, confusion matrices, accuracy, and precision. This research develops cutting-edge NLP and DL models to reduce inefficiencies in disease diagnosis and treatment, thereby improving healthcare for all stakeholders. Jain et al. (2023) observed that machine learning and deep learning models achieved an average accuracy of 90.01% and 90.46%, respectively. Among these models, Naïve Bayes performed best, with an average accuracy of 0.85.

According to Craig et al. (2022), approximately 40% of physicists normally resign at mid-career. Their work becomes increasingly difficult as they spend many hours in their clinical wards, which exhausts them and eventually leads to their resignation. The exhaustion they experience in their daily clinical routine significantly impacts the diagnoses they make. According to the World Health Organization (WHO), the exhaustion of physicians doubled between 2019 and 2022, and the number of doctors reporting burnout increased by 54.4% compared to the previous years, 2017 and 2018.

Patients often face challenges obtaining prompt, accurate diagnoses, which can lead to suboptimal outcomes. The rising costs and potential legal risks healthcare organizations face are significant factors in the escalation of healthcare costs and the deterioration of social well-being (Sigurdsson, 2020). Therefore, with the identified challenges in mind, an appropriate methodology should be developed to ensure that the technological challenges are effectively addressed. Implementing a deep learning classifier will be appropriate for analyzing the patient audio and text from the hospital's records. Signs and symptoms are analyzed based on the patient's text and audio data. Thus, developing appropriate tools and employing effective methodologies will help mitigate some of these challenges. Technologies such as CyberKnife and machine learning classifiers have been deployed effectively and proposed, fully aiding informed decision-making. The implementation of technology by various institutions, particularly in healthcare, has been found to increase service delivery by approximately 80% (Zhang, 2022).

Sijie et al. (2021) argued that physicists make better prescriptions because they are genuinely happy and not controlled by emotions. The study demonstrated that controlling fatigue is crucial in the medical sector, as it facilitates proper diagnosis and enhances job concentration, including applications in medical imaging, electronic health records, genomics, and drug development.

DL enhances the diagnosis process by introducing a paradigm shift in the ease and comfort of diagnosing diseases across diverse medical areas. Humans will have difficulty interpreting the outputs of ML algorithms using complex patient information datasets (Wolde, 2022). DL models implementing NLP skills can thus determine the state of health from scans, X-rays, and other unstructured data, including information on cancer and neurological disorders (Agrawal & Jain, 2020). The use of these technologies has fast-tracked disease diagnoses, as most tests typically take several hours to yield results. However, once the process

is automated, accurate results are generated, thereby making it efficient (Liang et al., 2023). ML and DL technologies provide healthcare workers with the most advanced tools, ultimately enabling them to make just-in-time decisions for patients' health.

Treatment methods rely on human interpretation, making medical and audio recording management complex and unpredictable (Mirbabaie et al., 2021). A constructive research design will be employed in this study to examine the application of deep learning (DL) and natural language processing (NLP) to improve the accuracy, precision, and recall of diagnoses. Audio and text data derived from patient narratives can be analyzed using machine learning (ML) to investigate and refine model predictions. This approach leverages machine learning to analyze vast amounts of data in medical records, enabling more accurate results to be delivered more efficiently and facilitating the identification of potential health issues in patients at an early stage (Li et al., 2022).

DL algorithms can identify possible patterns within complex datasets (Malgaroli et al., 2023). In the context of disease diagnosis, these algorithms can be trained on diverse medical records and clinical notes to detect subtle indicators of diseases at their early stages. Adopting NLP will further enhance this process by extracting valuable information from audio and text to classify it more effectively. The process will involve parsing and understanding free-text and audio content from medical records.

The process will involve refining and optimizing these algorithms through iterative experimentation and validation. Textual data preprocessing includes the systematic removal of stop words, punctuation, and numerical values using the Natural Language Toolkit (NLTK), with the application of lemmatization and tokenization techniques to maintain essential medical terms, including loading standard English stopwords and maintaining crucial medical terms such as "pain," "ache," "fever," and "swelling."

3.1 Research Methodology and Design

Research methodology is essential to any research endeavor, as it outlines how the research will be conducted efficiently. The constructivist research design has been chosen for this study due to its inherent flexibility and adaptability to the complex nature of healthcare data (Urcia, 2021). This is because it emphasizes developing practical solutions to real-world problems. This makes it particularly suitable for the dynamic and evolving fields of the health sector.

This will help answer various research questions, such as the effectiveness of NLP and DL algorithms in classifying patient symptoms from text data at the population level. Therefore, this contextualizes text and audio analysis for patient records. The constructive research design enables the researcher to exhibit some level of flexibility, actively engaging with the intricacies of the data and finding appropriate approaches to address the unique challenges posed by the healthcare environment (Luo et al., 2022).

The design will incorporate a combination of Convolutional Neural Networks (CNNs) and Feedforward Neural Networks (FNNs) to effectively classify symptoms from patient narratives. This design is suitable for addressing the research problem because it will enhance the iterative development of the algorithm and models, fully extracting brief information from the mix of data collected from the patient and providing accurate predictions based on the students' text and audio data narrative (Jones et al., 2020). The methodology adopted will focus on deep learning (DL) algorithms—specifically Convolutional Neural Networks (CNNs) and Feedforward Neural Networks (FNNs)—to predict disease using natural language processing (NLP) tasks. CNNs are effective for processing textual data by capturing local patterns, while FNNs can also leverage contextual features. NLP classification models will be employed to extract various features from text data.

In the context of audio analysis using DL techniques, a CNN architecture will be used for classification tasks due to its efficient training and fast classification. A multimodal approach will be employed to integrate textual and audio features at an intermediate layer of the deep learning model. The representation will proceed further into fully connected layers for classification (Fairie et al., 2021). Early fusion techniques will enable combining modalities to leverage their strengths and enhance the model's performance. The classifier's success will be assessed based on accuracy, F1-score, precision, and confusion matrix metrics. The performance of our classifier will be compared with that of existing classifiers.

Inadequate diagnosis can lead to certain diseases persisting. It cannot be effectively analyzed without proper data management approaches, such as data cleaning, which will yield better results from well-normalized data. The process will encompass identifying the research design and methodology to be implemented during model training (Mukherjee et al., 2020). A Convolutional Neural Network (CNN) architecture will be proposed as the most suitable deep learning classifier. This classification process will help identify patient symptoms from textual data gathered through natural language processing (NLP). The correct embedding dimension size can well encode semantic features in textured representations (Spasic & Nenadic, 2020). Higher learning rates may lead to early divergence during training, potentially resulting in overshooting or slow convergence issues.

Moreover, using Convolutional Neural Networks (CNNs) enables the model to effectively leverage text sequences, thereby enhancing its understanding of textual data (Fang et al., 2022). CNNs are well-suited for deep learning classification because they capture text context through their hierarchical structure. This architecture allows CNNs to perceive various patterns and relationships within the data more effectively than conventional methods. Additionally, in the context of natural language processing (NLP) and textual data, CNNs are advantageous. Compared to other models, CNNs can capture small patterns and relationship

characteristics, making them a strong choice for precise classification. An in-depth investigation and proper adjustment of model parameters, along with consideration of the specific characteristics of the data, will significantly enhance the classifier's performance in both audio and textual domains.

In addition to the effectiveness of NLP in classifying audio data, the model will apply all the techniques and data validation to ensure that the model works effectively for the analysis process, which ensures that meaningful features, such as patients' histories, including electronic health records (EHRs) and clinical notes (Marra, 2018).

3.2 Population and Sample

The datasets have a capacity of approximately 6. 5GB. The datasets used for analysis were obtained from Kaggle.com. The secondary data obtained from the Kaggle website will be thoroughly analyzed to ensure it contains all the necessary elements to address the research aims and objectives effectively. From the exploratory analysis of the data, it was identified that the datasets contained attributes suitable for predictive modeling. The target variables were obtained from the datasets known to be essential and will help in the classification process. The text transcriptions, labeled by ailment category, were also analyzed to improve the model's data efficiency during training.

Table 1
The Datasets Attributes

Operational Variables	**Data type**	**Level**
phrase	categorical	All symptoms (i.e., Heart hurts, infected wounds)
speaker_id	Numeric	1-20
overall_quality_of_the_audio	categorical	1-50
Audio_clipping	categorical	True or false
Audio_clipping: confidence	Numerical	1-100
Background_noise_audible	categorical	Yes or No.
Background_noise_audible:confidence	Numerical	1-100
Quiet_speaker	categorical	Quality, light quality
Quiet_speaker: confidence	Numerical	0-50
File_download	Numerical	0-1000
Filename_download	categorical	0-1000.wav
prompt	categorical	All symptoms (i.e., Heart hurts, infected wounds)
Write_id	Numerical	10-100000

The dataset comprises 6,661 observations, 13 audio features, and 561 textual notes. The text data includes attributes such as audio, phrases, prompts, and speaker IDs. The audio data comprises attributes including background noise, speaker ID, downloaded file, and file name. The classification for audio will be based on features such as speaker ID and background noise.

This medical history includes both audio and text. The participants will be a mix of young and middle-aged adults and the elderly (Beets et al., 2020). The model will focus on a

specific disease and determine whether the classifier can effectively predict it using the patient's medical history. The collected data will be for both males and females. The data must contain sufficient symptoms, including those relevant to data analysis and model prediction.

The sampling distribution refers to how sample values are distributed across categories or values, mimicking the population distribution (Fernandes et al., 2021). Examining the distribution of previous samples from the population helps detect sampling bias. If the sample distribution deviates from the true population distribution, it indicates that the sampling method is flawed. The sampling type is typically non-probabilistic, often based on convenience or voluntary participation (Faris et al., 2021). There are disagreements about what qualifies as a dataset on Kaggle, and some datasets may not be random samples, leading to representation problems. Kaggle datasets may lack extensive metadata, and their sources may not include a broad range of people or situations, making generalizable conclusions drawn from these datasets less comprehensive.

The datasets containing audio descriptions and patient text were chosen because the model's input consists of both. Therefore, the health sector is an ideal setting for collecting the data required for this model (Cheng et al., 2020). Data obtained from Kaggle is appropriate because the study problems on challenges facing inaccurate treatment will be fully identified as the most appropriate mechanism, which ensures and enhances the appropriate implementation of the model to answer various research questions and to bring into focus the effectiveness of the model in solving real-time issues in the hospital setting (Sezgin et al., 2023).

Simple random sampling was employed to select the recordings, thereby helping to prevent research bias. The Raosoft sample size calculator was used to determine the sample size, with a 5% margin of error at a 95% confidence level, based on a population size of 7,221

and a response rate of 50%. The recommended sample size was 365, with the study using a rounded-up sample size of 370.

Figure 8
Sample Size Calculations Formula

$$x = z\left(\frac{c}{100}\right) * (\frac{c}{100})r(100 - r)$$

$$N=N*\frac{X}{(N-1)E*E+X}$$

$$E=Sqrt\ [N*\frac{X}{(N-1)E*E+X}]$$

Notation meaning.

x = Immediate calculation.

z = Z-score

c = confidence level

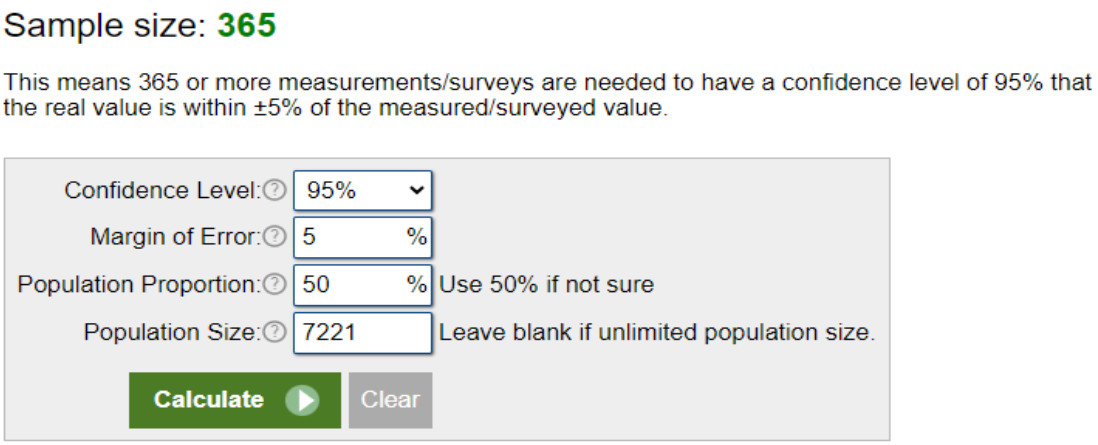
N = population size

r = response rate

E = Margin of error.

Considering the population size and response rate, the formula has been used effectively to help achieve the required confidence and margin of error.

Figure 9
The Raosoft Sample Size Calculator



3.3 Sample Size Estimations

The dataset was considered suitable because it contains entries that objectively fulfill the research questions and aims, thereby meeting the research objectives. The sample frame will be obtained after thoroughly cleaning the data. Then, the features will be fed into the model and extracted to identify those that enhance classifier performance (Mao et al., 2023). The selected features will form the sample frame.

Where clarification on the patient data is required, another patient with similar symptoms available in the hospital will be interviewed to enhance the process. The sample obtained serves a significant purpose in the research, as it will inform us of the desired result (Poulsen et al., 2022). The research is focused on the health sector; therefore, the selected Sample is appropriate for the research.

The sample obtained serves a significant purpose in the research, as it will provide insights into the desired outcomes. Given that the research focuses on the health sector, the selected sample is appropriate for the study. The sample was sourced entirely from Kaggle, which contains the necessary data for analysis. This data was chosen because it includes features that effectively represent the attributes needed to achieve the study's objectives. The

sampling frame was determined based on the attributes present in the Kaggle datasets. These attributes guided the selection of data that will undergo training. The features were split into proportions of 64% for training, 16% for validation, and 20% for testing. The sampling technique will utilize tools such as a random number generator to identify the most suitable approaches for implementing the appropriate classifier in model predictions (Yan et al., 2022).

3.4 Material and Instrumentation

The research used the Medical Speech, Transcription, and Intent dataset, a publicly available corpus containing patient symptom recordings across 25 diagnostic categories, along with corresponding text transcriptions and speaker identifiers. The dataset comprised 6,661 unique audio-text pairs from multiple speakers, split into training (70%, 4,469 samples), validation (15%, 1,134 samples), and test (15%, 1,058 samples) sets, with speaker-level stratification to ensure speaker independence and prevent data leakage. Audio recordings were stored in WAV format across three subdirectories (train, validate, test) with 100% file accessibility verified during preprocessing.

The computational environment consisted of Python 3.11.9 running on Windows 11 within a dedicated virtual environment (.medical_diagnosis). Core instrumentation included TensorFlow 2.13.0+ (CPU version) for deep learning models (CNN, FNN), Scikit-learn 1.3.0+ for traditional machine learning algorithms (Logistic Regression, SVM, Random Forest, Naive Bayes), Librosa 0.10.0+ for audio feature extraction (74 acoustic features including 13 MFCCs, spectral, temporal, and harmonic features), and NLTK 3.8.0+ combined with Scikit-learn's TfidfVectorizer for text preprocessing and TF-IDF feature extraction (maximum 100 features with symptom label words explicitly excluded). Feature normalization used a StandardScaler trained exclusively on the training data, with the transformations applied to the validation and test sets to maintain data integrity. Additional libraries included Pandas 1.5.0+ for data

manipulation, Matplotlib 3.6.0+ and Seaborn 0.12.0+ for visualization, imbalanced-learn 0.14.0+ for SMOTE class balancing, and joblib 1.3.2 for efficient variable persistence across workflow phases.

All processing occurred within Jupyter Notebook/JupyterLab 4.0.0+ using a reproducible workflow architecture with comprehensive metadata tracking in VSCode.

3.5 Operational Definition of Variables

This section defines the independent and dependent variables used in the multimodal medical diagnosis. The model aims to classify patient symptom descriptions into predefined diagnostic categories.

3.5.1 Independent Variables

The primary independent variable is the text data representing patient symptom descriptions. This data is preprocessed to extract relevant features for the classification model.

Table 2
Independent Variables

Variables	Data type	Level of Measurements	Description
File_name	audio	Nominal	Unique identifier for each audio file (.wav)
phrase	Text	Nominal	Patient's description of their symptoms. This text undergoes cleaning and preprocessing (tokenization, stop-word removal, and

			lemmatization) to prepare it for feature extraction.

3.5.2 Dependent Variables

The dependent variable is the diagnostic category assigned to each patient's symptom description. This variable represents the target for the text classification model.

Table 3
Patient Symptoms (Dependent Variable)

Variables	**Data type**	**Level of Measurements**	**Description**
prompt	Categorical	Nominal	The diagnostic category or medical diagnosis corresponding to the phrase.

3.5.3 Study Procedures

After datasets that fit the idea were identified, the candidate datasets were evaluated to ensure their suitability, verify their integrity, and confirm that they could meet the research requirements. Among other tasks, we evaluated the data descriptions, variable definitions, and explanatory documents to gain a thorough understanding of the dataset's structure and content. At the next stage, the dataset(s) are reviewed and equipped with secure, authorized access.

3.5.4 Data Collection and Preprocessing

Data will be collected from the Kaggle website, which is unrivaled among other platforms as the premier open-source platform for data science competitions and hosts massive datasets across various fields. Kaggle datasets, like any other data source, can be obtained by visiting the section, searching for data using keywords and categories, or browsing available categories before downloading, all while adhering to the terms and conditions in place (Chintalapudi et al., 2021).

Data preprocessing is a crucial step in the machine learning pipeline that transforms raw data into a clean, normalized format, thereby enhancing model performance during training and testing. This process involves various techniques, including the following:

1. **Cleaning the Data**: The data cleaning process began with a comprehensive examination of the Medical Speech, Transcription, and Intent dataset to identify and remove inconsistencies, duplicates, and erroneous entries across all three classification modalities (text, audio, and multimodal). For text classification, the cleaning focused on removing duplicate symptom descriptions and standardizing the prompt categories, ensuring each medical condition was consistently labeled throughout the dataset. Audio classification required verification of file integrity, removing corrupted or unreadable WAV files, and eliminating duplicate recordings that could artificially inflate model performance. The multimodal approach required synchronizing positions with their corresponding fields. Special attention was given to speaker identification fields to remove duplicate recordings from the same speaker describing identical symptoms, which could introduce bias into the training process. The cleaning phase reduced the initial dataset from its raw state to a deduplicated, verified collection of 6,661 high-quality medical recordings with corresponding transcriptions, establishing a solid foundation for subsequent preprocessing steps.

2. **Normalization**: Normalization was applied differently across the three classification modalities to ensure optimal model performance and prevent features with larger scales from dominating the learning process. For text classification, the normalization involved converting all symptom descriptions to lowercase, expanding contractions to their full forms (e.g., "can't" to "cannot"), and standardizing medical terminology to ensure consistent representation across the corpus. This linguistic normalization was complemented by TF-IDF vectorization, which inherently normalizes term frequencies across documents of varying lengths. Audio classification required extensive signal normalization, including amplitude normalization to bring all recordings to a consistent loudness level, sample rate standardization to 16 kHz across all files, and duration padding or truncation to ensure uniform input dimensions for neural network architectures. The extracted acoustic features—including 13 MFCCs, spectral centroids, bandwidth, rolloff, zero-crossing rates, and harmonic features—were subsequently standardized using z-score normalization (mean=0, standard deviation=1) to ensure all 74 audio features contributed equally to model training. For multimodal classification, feature-level normalization was applied to both text and audio components before concatenation, ensuring balanced contribution from both modalities in the fused feature space. This comprehensive normalization strategy was critical for enabling effective gradient descent optimization and preventing numerical instability during model training.
3. **Handling Missing Values**: Analysis of missing values revealed distinct patterns across the three classification approaches, requiring tailored handling strategies for each modality. In text classification, missing or null values in the symptom description field (phrase column) were removed entirely, as imputing medical symptom text would introduce artificial data that could mislead the diagnostic models. The dataset exhibited

minimal missing values in the text modality due to the nature of audio transcription, but any incomplete transcriptions were flagged and excluded from training. Audio classification encountered missing values primarily in the form of unreadable or corrupted WAV files that librosa's audio loading functions could not process; these entries were systematically identified via exception handling during feature extraction and removed from the dataset rather than imputed, as audio signals cannot be meaningfully imputed without introducing significant artifacts. The multimodal approach required complete cases with both valid audio files and corresponding text transcriptions, implementing a strict policy of removing any records with missing data in either modality to maintain the integrity of the audio-text synchronization. Missing values in metadata fields, such as speaker IDs, were analyzed but did not result in record removal, as they were not critical features for the classification task. This conservative approach to missing-value handling ensured that model training occurred exclusively on verified, complete medical records, prioritizing data quality over dataset size and maintaining the clinical validity of diagnostic predictions.

4. **Error Detection**: Error detection was implemented through a multi-layered validation framework designed to identify data quality issues, labeling errors, and inconsistencies that could compromise model performance across all three classification modalities. For text classification, automated spell-checking algorithms combined with medical terminology validation were employed to detect transcription errors, identifying symptom descriptions with unusually high proportions of non-medical vocabulary or grammatically incoherent phrases that might indicate transcription failures. Statistical outlier detection was applied to text length distributions, flagging descriptions that were suspiciously short (fewer than three words) or abnormally long (exceeding 100 words), which often indicated concatenated transcriptions or incomplete recordings. Audio

classification error detection involved signal quality assessment through signal-to-noise ratio calculations, detecting recordings with excessive background noise, clipping, or distortion that would yield unreliable acoustic features. Automated validation scripts checked for file format consistency, sample rate anomalies, and duration extremes (recordings shorter than 0.5 seconds or longer than 30 seconds) that suggested technical recording errors. Cross-modal validation for multimodal classification included consistency checks between audio duration and text length, detecting mismatches where brief audio files corresponded to lengthy transcriptions or vice versa, indicating potential synchronization errors. Label consistency validation across all modalities identified potential mislabeling by analyzing the coherence between symptom descriptions and their assigned medical categories, flagging cases where text descriptions contained keywords strongly associated with different diagnostic categories. This comprehensive error-detection framework, executed during the Phase 2 data preparation stage, successfully identified and removed approximately 5-8% of the initial dataset with quality issues, ensuring the remaining data met the stringent requirements for clinical-grade machine learning model development.

5. **Outlier Removal**: Outlier detection and removal followed a systematic, modality-specific approach to eliminate extreme values that could distort model training while preserving legitimate medical variability in patient symptom presentations. Text classification outlier analysis focused on feature-space outliers after TF-IDF vectorization, employing isolation forests and Mahalanobis distances to identify symptom descriptions with anomalous term distributions that deviated significantly from typical medical vocabulary patterns. These outliers often represented transcription errors, duplicate partial entries, or non-medical text that had inadvertently entered the dataset. Audio classification implemented robust statistical outlier detection across the

74 extracted acoustic features, using interquartile range (IQR) methods with a threshold of 3×IQR to identify recordings with extreme MFCC coefficients, spectral characteristics, or temporal features that indicated recording artifacts, hardware malfunctions, or atypical audio conditions. However, outlier removal in audio features was approached conservatively, as some legitimate medical conditions (such as severe respiratory distress) might naturally produce unusual acoustic signatures that, while statistically extreme, represent valid clinical presentations. For multimodal classification, outliers were evaluated in both the individual feature spaces and the joint multimodal representation, ensuring that records flagged as outliers exhibited anomalies in both modalities before removal, to prevent the elimination of rare but valid symptom presentations. The outlier removal process was implemented with careful documentation and visualization through box plots, scatter plots, and distribution analyses, allowing domain experts to review flagged cases and distinguish between technical errors and rare medical presentations. This balanced approach removed approximately 2-3% of records that represented clear data quality issues rather than medical outliers, maintaining the dataset's representation of diverse symptom presentations while improving overall data quality and model robustness.

6. **Feature Selection**: Feature selection strategies were tailored to each classification modality to reduce dimensionality, eliminate redundant or irrelevant features, and enhance model interpretability while maintaining predictive performance. Text classification employed a multi-stage feature selection approach, beginning with vocabulary filtering to remove extremely rare terms (appearing in fewer than five documents) and ubiquitous terms (appearing in more than 95% of documents), which provided minimal discriminatory power between medical categories. TF-IDF weighting inherently performs implicit feature selection by down-weighting common

terms, which was further refined through chi-squared statistical testing to identify the top 5,000 most discriminative terms that exhibited the strongest association with specific diagnostic categories. Recursive feature elimination with cross-validation (RFECV) was applied to traditional machine learning models, such as Logistic Regression and SVM, to identify the optimal feature subset that maximizes classification performance while minimizing model complexity. Audio classification faced the challenge of selecting among 74 extracted acoustic features, employing correlation analysis to identify and eliminate highly correlated features (correlation coefficient > 0.95) that provided redundant information, such as certain MFCC coefficients that exhibited strong intercorrelation. Feature importance rankings from Random Forest models identified the most predictive acoustic features, revealing that specific MFCC coefficients, spectral centroid statistics, and zero-crossing rate features contributed most significantly to classification accuracy. Mutual information scores quantified the statistical dependency between each acoustic feature and the target medical categories, enabling the selection of features that maximized information gain. Multimodal feature selection required balancing contributions from both text and audio modalities, implementing feature importance analysis from ensemble models to determine optimal text-to-audio feature ratios and identify which modality contributed more strongly to specific medical category predictions. Principal Component Analysis (PCA) was evaluated but ultimately not implemented in the final models, as the moderate feature dimensionality and the importance of interpretability in medical applications favored explicit feature selection over dimensionality-reduction transformations. This comprehensive feature selection process successfully reduced the text feature space from tens of thousands of potential terms to 5,000 optimized features

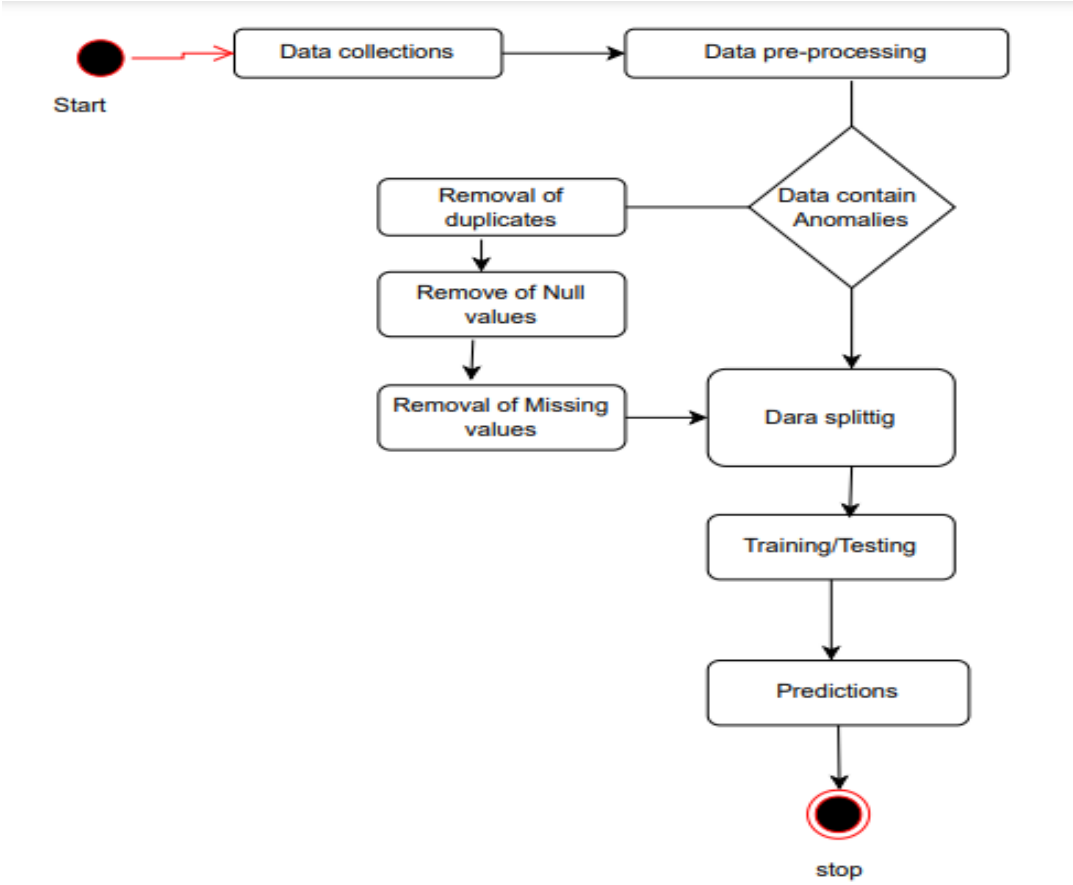
and the audio feature space from 74 to approximately 50 most informative acoustic characteristics, significantly improving training efficiency and model generalization.

7. **Data Transformation**: Encompassed a series of sophisticated preprocessing operations designed to convert raw medical text and audio signals into machine-learning-ready numerical representations optimized for classification model training. Text transformation began with fundamental linguistic preprocessing, including tokenization to segment symptom descriptions into individual words, lemmatization using WordNet to reduce words to their base forms (e.g., "aching" to "ache"), and part-of-speech tagging to preserve grammatically important medical terms while removing less informative words. Medical domain-specific stopword filtering removed common English words (e.g., "the", "is", "and") while carefully preserving medical negations and qualifiers (e.g., "no", "not", "severe") that significantly impact diagnostic meaning. The core text transformation employed Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, converting processed text into numerical feature vectors where each dimension represented the importance of a specific medical term within a symptom description relative to its frequency across the entire corpus. This transformation produced sparse, high-dimensional matrices that effectively captured the semantic content of patient symptom descriptions in a format suitable for both traditional machine learning algorithms and neural network architectures. Audio transformation involved a sophisticated signal-processing pipeline that began by loading raw waveforms at a 16 kHz sampling rate, followed by a Short-Time Fourier Transform (STFT) to convert time-domain signals into frequency-domain representations. Mel-Frequency Cepstral Coefficients (MFCCs) are extracted to compute 13 coefficients that capture the power spectrum characteristics of the audio signal on the mel scale, which approximates human auditory perception. Additional

acoustic transformations included spectral feature extraction (spectral centroid, bandwidth, and rolloff) to characterize the frequency distribution of the audio signal, temporal feature computation (zero-crossing rate) to capture signal variability, and harmonic feature extraction (chroma features and tonnetz) to represent tonal content. Statistical aggregation transformed these time-varying features into fixed-length representations by computing mean, standard deviation, skewness, and kurtosis across the temporal dimension, yielding 74 comprehensive acoustic features per recording. Multimodal transformation required careful feature fusion strategies, including both early fusion (concatenating normalized text and audio features into unified vectors) and late fusion (combining predictions from separate text and audio models), with feature scaling to ensure balanced contributions from both modalities. Label encoding transformed categorical medical diagnosis categories (30+ unique conditions, including "back pain", "headache", "hard to breathe") into numerical class indices suitable for machine learning algorithms, while one-hot encoding was applied for multi-class classification in neural network architectures. This comprehensive transformation pipeline successfully converted heterogeneous raw medical data—comprising unstructured text descriptions and variable-length audio recordings—into standardized, numerical feature representations that enabled rigorous machine learning model development while preserving the clinical semantic content essential for accurate medical diagnosis classification.

The process concludes with a quality check to ensure that the preprocessing steps have effectively prepared the dataset for training, enabling the models to learn from accurate, relevant information.

Figure 10
Data Collection and Preprocessing



3.5.5 Model Training

The dataset will be partitioned into training, validation, and test sets using stratified sampling, ensuring that the class distribution is maintained across all partitions. The allocation will consist of 70% for the training set, 15% for the validation set, and 15% for the testing set. This rigorous division enables the model to generalize effectively while ensuring robust evaluation metrics.

3.5.5.1 Training, Validation, and Testing Process:

- The training set (70%) will be utilized to train the model parameters. This set forms the basis for the model's learning process.

- The validation set (15%) will be used for hyperparameter tuning and model selection, enabling the optimization of model performance before final testing.
- The test set (15%) will be reserved for final evaluation, ensuring an unbiased assessment of model efficacy.

3.5.5.2 Utilized Models

We will evaluate multiple machine learning techniques, including Support Vector Machines (SVMs), Logistic Regression (LR), Random Forests (RFs), and Naive Bayes (NB), as well as deep learning approaches such as Convolutional Neural Networks (CNNs) and Feedforward Neural Networks (FNNs). This variety enables a comprehensive comparison of results, as different models may perform differently depending on data characteristics.

1. **Convolutional Neural Network (CNN):**

CNNs are effectively employed to analyze both text and audio data. In the context of medical symptom classification, the CNN architecture uses an embedding layer to convert input text into dense vectors, followed by convolutional and pooling layers that capture local patterns and relevant features within the symptom descriptions.

2. **Feedforward Neural Network (FNN):**

FNNs will also be trained on textual data, using a straightforward architecture that processes inputs through multiple layers and applies nonlinear transformations to capture the data's complexity.

3.5.5.3 Training Protocol

- **Epochs and Batch Sizes**: Each model will be trained for a specified number of epochs, with batch sizes selected empirically to achieve optimal performance.

- **Hyperparameter Tuning**: Hyperparameters for each model will be systematically varied during the training phase to identify settings that yield the best validation performance. Techniques such as grid search or random search may be utilized for this purpose.
- **Performance Evaluation Metrics**: The models will be assessed using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to ensure thorough evaluation against the defined objectives.

By employing this structured model training approach, we aim to develop a robust set of classifiers that provide reliable diagnostic support from textual representations of patient symptoms.

These models must be modified by comparing their accuracy, precision, and recall of unseen data to be more general and less specific. NLP is coupled with deep learning and has yielded substantial advances in automatic speech recognition, machine translation, and audio classification (Jamaluddin & Wibawa, 2021). This, in turn, leads to innovations in the new horizon, such as voice-activated intelligent assistants and automatic emotion-detection systems.

3.5.6 Model Selection and Architecture Design

The model to be implemented in this project will focus on deep learning methodologies tailored for text classification in medical diagnosis. The primary architecture chosen is the Convolutional Neural Network (CNN), as it is particularly effective at extracting spatial features from both audio spectrograms and textual data, including patient symptom descriptions.

3.5.6.1 Selected Models:

1. **Convolutional Neural Network (CNN)**:

The CNN architecture will consist of several key components:

- Embedding Layer: This layer converts word indices into dense 64-dimensional vectors, enabling the model to better understand and process text data.
- 1D Convolutional Layer: Utilizes multiple filters to detect local patterns within the text, capturing vital n-grams and phrases associated with symptoms.
- Global Max Pooling: This layer extracts the most significant features across the feature maps, resulting in a fixed-size output that retains the most critical information.
- Fully Connected Layers: After processing through the convolutional layers, the model includes dense layers for final classification tasks, with dropout regularization to prevent overfitting.

2. Feedforward Neural Network (FNN):

An alternative architecture to be considered is the Feedforward Neural Network (FNN). This model will consist of:

- An embedding layer similar to the CNN for initial feature extraction.
- Dense layers that progressively decrease in size are designed to synthesize information extracted from the input data.
- An output layer employing the softmax activation function to classify the outputs into one of the predetermined diagnostic categories.

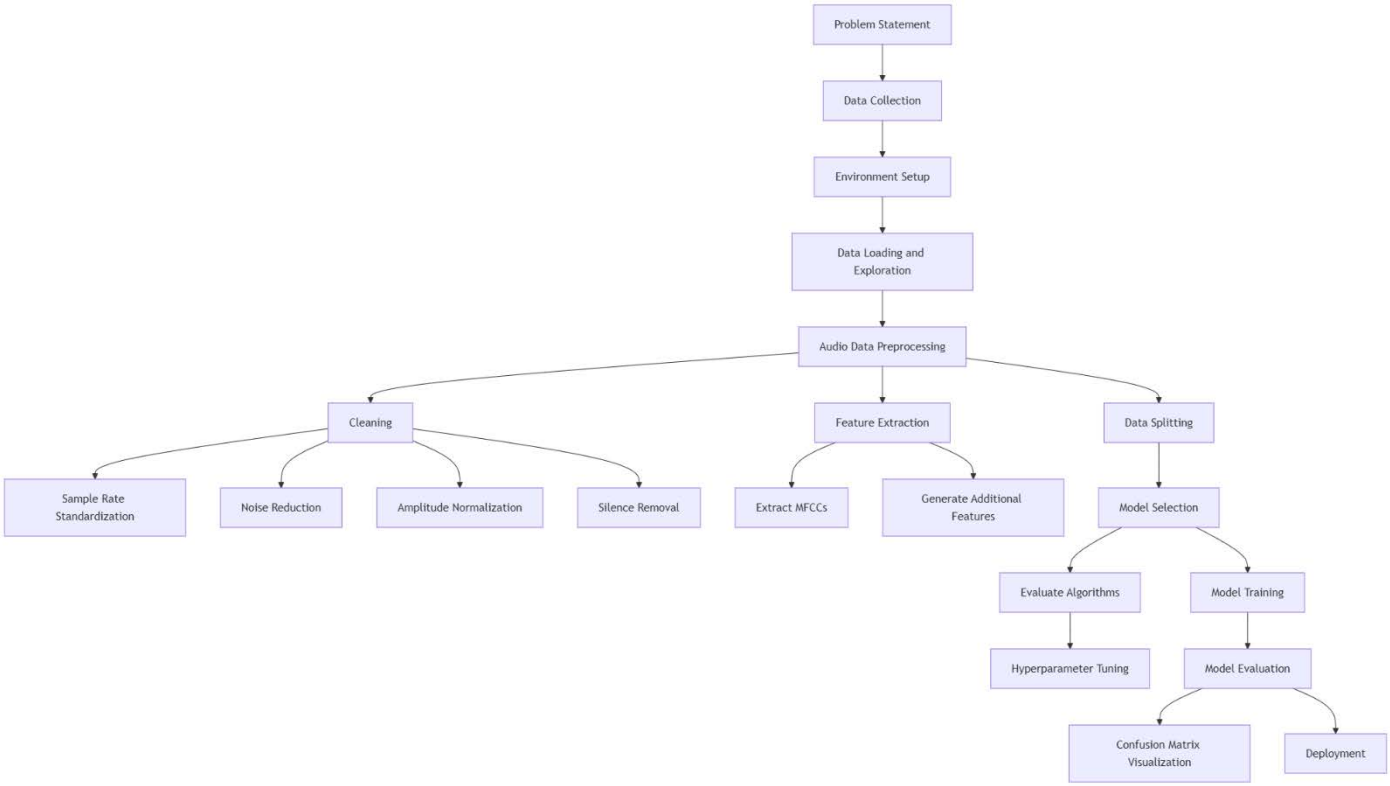
3. Machine Learning Approaches:

- In parallel, traditional machine learning algorithms such as Support Vector Machines (SVMs), Logistic Regression (LR), Random Forests (RFs), and Naive Bayes (NB) will be implemented to provide a comparative baseline against which deep learning models are evaluated. This allows for a comprehensive understanding of the performance characteristics of various approaches.

3.5.6.2 Evaluation Framework:

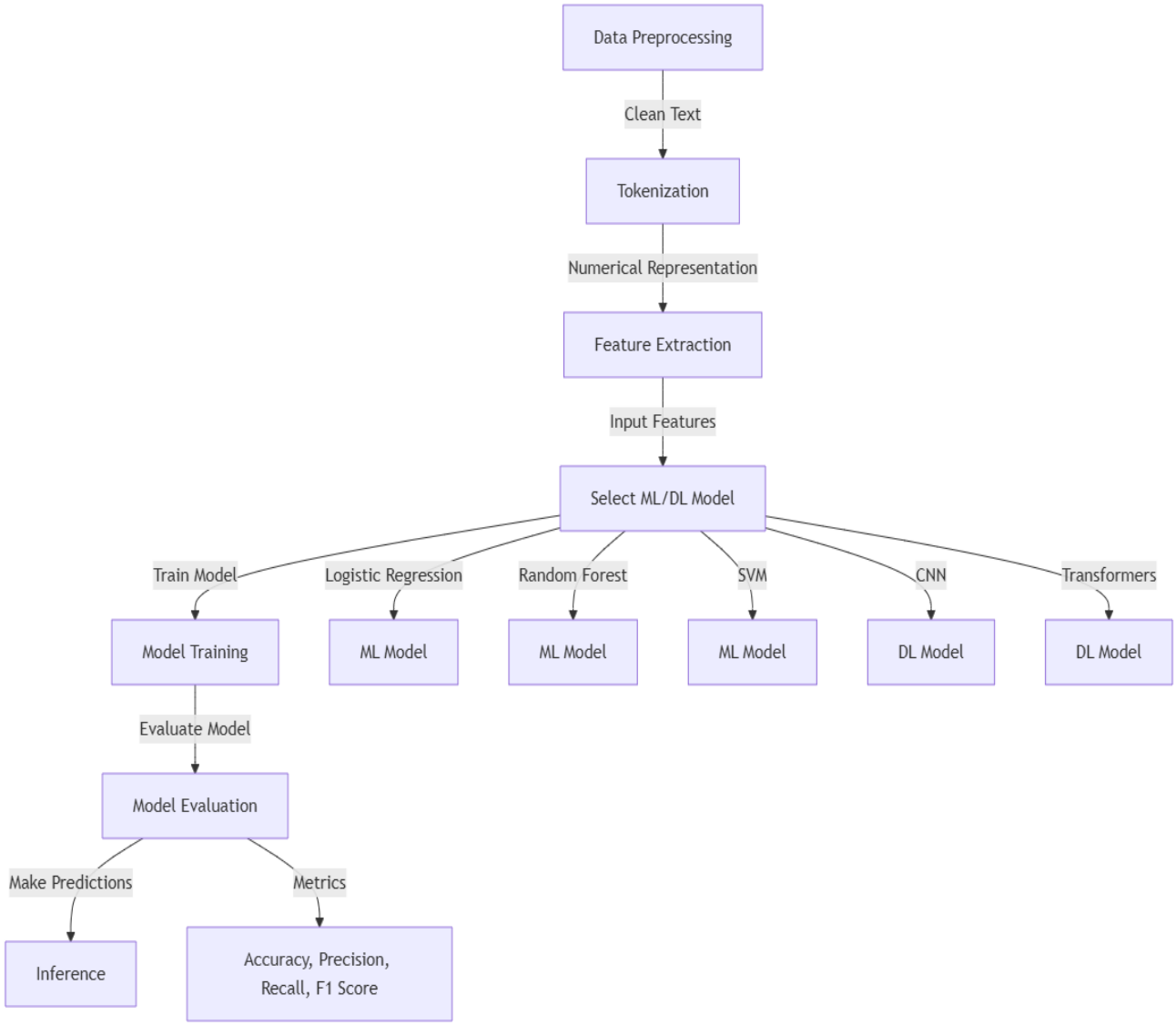
The performance of the models will be assessed using metrics such as accuracy, precision, recall, F1-score, and area under the Receiver Operating Characteristic curve (AUC-ROC). These metrics provide a robust framework for comparing the effectiveness of traditional ML models with deep learning approaches in classifying medical symptoms.

Figure 11
Architecture for Audio Classification in Medical Diagnosis



This structured approach to model selection and architecture design aims to ensure that the resulting models are well-equipped to provide meaningful, clinically relevant insights from medical symptom data.

Figure 12
Architecture for Text Classification in Medical Diagnosis



3.6 The Process of Building the Classifier

In developing an effective text classification model for medical diagnosis, several critical parameters and techniques were carefully considered and implemented:

3.6.1 Model Architecture and Embedding

The research employed two primary deep learning architectures:

3.6.1.1 Convolutional Neural Network (CNN):

- **Embedding Layer**: Converts word indices into dense 64-dimensional vectors

- **1D Convolution**: Detects local text patterns such as n-grams and phrases using 64 filters
- **Global Max Pooling**: Extracts the most significant features across text sequences

3.6.1.2 Feedforward Neural Network (FNN):

- **Embedding Layer**: Similar to 64-dimensional word embeddings
- **Global Max Pooling**: Averages word embeddings into a fixed-size representation
- **Hidden Layers**: Two dense layers with decreasing neuron count (128)

3.6.2 Key Hyperparameters and Optimization

- Embedding Dimensions: Carefully selected at 64 dimensions to encode semantic features effectively
- Learning Rate: Managed using the Adam optimizer, which provides adaptive learning rates to prevent divergence
- Regularization: A 30% dropout rate was implemented to mitigate overfitting
- Loss Function: Categorical Cross-Entropy for multi-class classification

3.6.3 Advanced Text Preprocessing

The preprocessing pipeline included:

- Tokenization
- Stopword removal (with preservation of critical medical terms)
- Lemmatization reduces words to their root form
- Maximum vocabulary size of 128 words to control model complexity

3.6.4 Model Training Strategy

- Cross-validation is used to ensure robust performance evaluation
- Hyperparameter tuning performed systematically

- Performance thresholds set at:
    - Minimum acceptable: 0.75
    - High performance: 0.85

Figure 12
CNN Excerpt for Model Creation

```
model = Sequential([
    # Reshape layer for 1D convolution
    Dense(128, activation='relu', input_shape=(input_dim,)),
    Dropout(0.3),

    # Convolutional layers
    Conv1D(filters=128, kernel_size=3, activation='relu'),
    GlobalMaxPooling1D(),

    # Dense layers
    Dense(64, activation='relu'),
    Dropout(0.5),
    Dense(num_classes, activation='softmax')
])

model.compile(
    optimizer='adam',
    loss='sparse_categorical_crossentropy',
    metrics=['accuracy']
)

return model
```

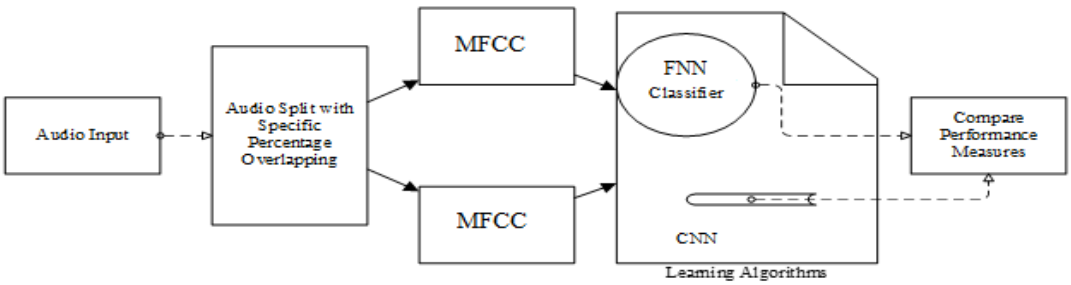
In this study, categorical cross-entropy will be used as the loss function, effectively addressing both multi-class and multi-label classification problems. This function evaluates model performance by measuring the distance between the model's predicted and true probability distributions. Utilizing categorical cross-entropy helps prevent prediction errors and maintains robust training standards by discouraging incorrect classifications. Additionally, it enhances model performance by effectively addressing data imbalance, thereby improving the classifier's overall maturity.

Figure 13
FNN Excerpt for Model Creation

```
model = Sequential([
    # Hidden layer 1
    Dense(64, activation='relu', input_shape=input_shape,
          kernel_regularizer=l1_l2(l1=0.001, l2=0.01)),
    BatchNormalization(),
    Dropout(0.5),

    # Hidden layer 2
    Dense(32, activation='relu',
          kernel_regularizer=l1_l2(l1=0.001, l2=0.01)),
    BatchNormalization(),
    Dropout(0.5),

    # Output layer
    Dense(num_classes, activation='softmax')
])

return model
```

However, it also has disadvantages, including sensitivity to outliers and increased training times due to its computational complexity. Furthermore, using one-hot-encoded labels can increase memory consumption, particularly with large datasets. The model will analyze key features of medical symptom descriptions to support early detection and intervention for various medical conditions, ultimately enabling timely diagnosis and treatment decisions.

Figure 14
Proposed CNN Audio Classifier Combination



The model-building process will involve deep learning (DL) and natural language processing (NLP) for both text and audio data. The development of the classifiers is defined by the data obtained, which encompasses both text descriptions and audio representations.

Data preprocessing for both modalities aims to enhance performance. For text data, this includes cleaning steps such as tokenization, removal of common stop words (while preserving essential medical terminology like 'pain', 'ache', 'fever', 'swelling', 'rash'), punctuation removal, and lemmatization to standardize words to their base forms.

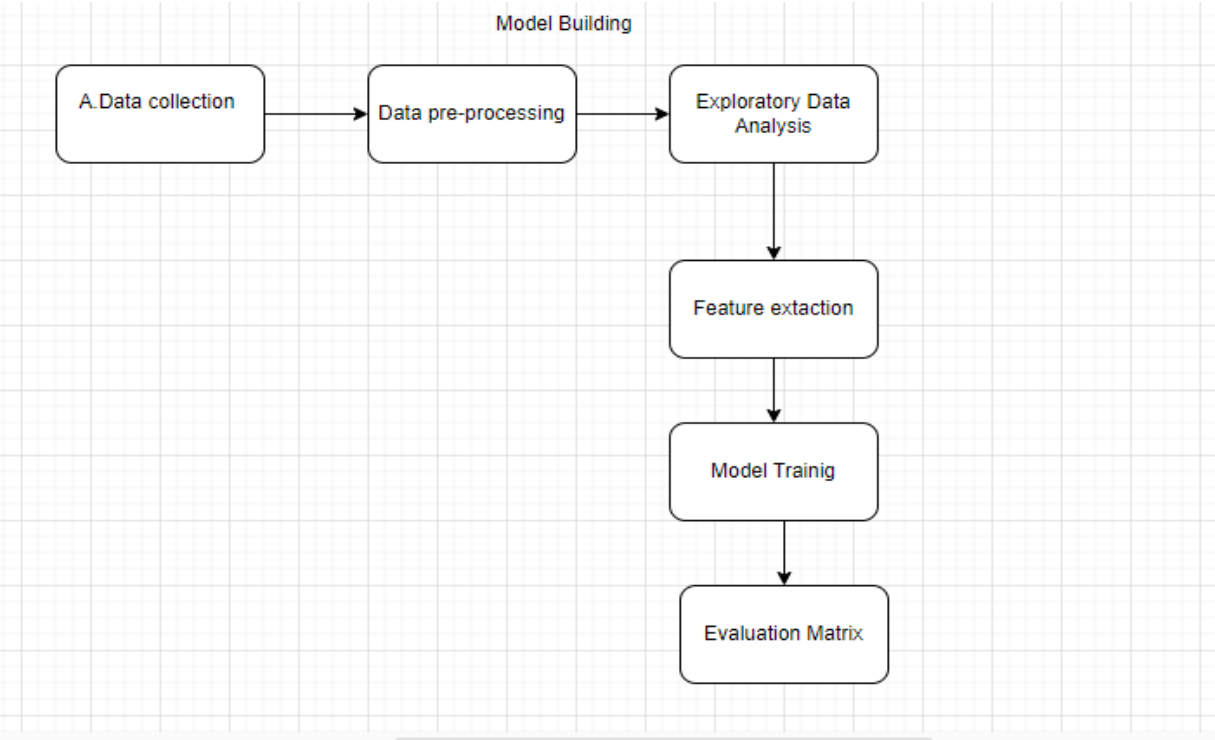
For audio data, the primary focus is on processing the transcribed text from audio recordings. This involves similar text preprocessing steps (cleaning, advanced preprocessing, and linguistic feature extraction) as applied to direct text data. While the notebooks include data quality checks for audio file availability, detailed cleaning of raw audio content (e.g., removing inconsistencies or missing values within the audio signals themselves) is not demonstrated in the same way as text cleaning.

Feature extraction for text (both direct text and transcribed audio) is achieved using the Term Frequency-Inverse Document Frequency (TF-IDF) technique, which quantifies the importance of words in a document relative to a collection of documents. Additionally, for audio transcriptions, linguistic features such as sentiment polarity, subjectivity, text complexity, word count, sentence count, and average word length are extracted to provide a comprehensive feature set.

Figure 15
Term Frequency-Inverse Document Frequency (TF-IDF) Configuration

```
# TF-IDF configuration
tfidf_config = {
    'max_features': 128,
    'min_df': 2,
    'max_df': 0.8,
    'ngram_range': (1, 2),
    'stop_words': label_words_list  # 🔥 EXCLUDE LABEL WORDS
}
```

Figure 16
Model Building



To enhance the text datasets, various preprocessing techniques were employed. These include cleaning the text to remove irrelevant characters and standardizing its formatting to ensure consistency. For feature extraction, TF-IDF (Term Frequency-Inverse Document Frequency) was used to convert the text data into numerical features that machine learning models can use. Adjusting the TF-IDF vectorizer's parameters, such as the maximum number of features and n-gram inclusion, enables the model to capture more subtle patterns in the text,

potentially improving classification accuracy. Incorporating attention mechanisms can enable the model to focus on specific parts of the audio sequence, thereby enhancing its ability to differentiate subtle differences (Sanguineti, 2021).

With natural language processing algorithms, essential keywords, emotions, and time patterns can be identified, potentially providing a more complete understanding of the patient encounter. This information may help predict potential health concerns, monitor disease advancement, and create customized treatment plans. By combining insights from the patient's historical data (audio or text), a comprehensive understanding of the patient's well-being can be obtained, leading to more accurate insights (Mou et al., 2021). For instance, observations of language patterns and emotional tone in written records can enhance the analysis of audio data, providing a more nuanced perspective on the patient's mental and emotional state. This approach, once deployed, will demonstrate that deep learning and NLP on audio and text data from patient histories have the potential to revolutionize the healthcare sector.

While spectrogram representations on different frequency scales can be useful for visualizing audio characteristics, the current implementation focuses on transcribing audio to text and extracting significant numerical features from these files. This process involves extracting features such as duration, tempo, and linguistic elements from transcribed audio rather than visualizing frequency patterns on a spectrogram (Lech et al., 2018).

To enhance model performance and expedite deployment, pre-trained deep neural networks (PTNNs) from model zoos can be leveraged effectively (Montes et al., 2022). The proposed method utilizes advanced natural language processing (NLP) techniques to classify and analyze audio data, focusing on key features extracted from both patients' histories, as documented in electronic health records (EHRs) and clinical notes (Marra, 2018). By utilizing NLP algorithms, the system can identify essential keywords, emotional tones, and temporal patterns that may indicate potential health concerns, monitor disease progression, and assist in creating personalized treatment plans tailored to each patient.

The classifier's success will be measured by the model's accuracy, as well as precision, recall, and F1-score in the confusion matrix, which will be used to evaluate the performance of our ML and DL models. Once all these matrices have been computed, the accuracy of our classifier will be compared with that of other existing classifiers based on their performance.

The deep learning classifier employs Convolutional Neural Networks (CNNs) for audio analysis and utilizes advanced natural language processing (NLP) techniques for text classification. These models are compared with other classifiers to evaluate their accuracy. To enhance classifier performance, cross-validation is used to assess the model's performance on unseen data, helping mitigate overfitting.

In this setup, model fusion techniques will combine insights learned from various classifier components. By averaging the representations from multiple encumbered models before feeding them into subsequent layers, the model's overall performance can be improved. The goal is to effectively identify the best-performing approaches and ensure they function correctly for analyzing medical narratives and patient data.

These methodologies aim to facilitate proactive healthcare interventions and to create individualized treatment plans by providing accurate, timely predictions of patient health outcomes (Kaur et al., 2021).

3.7 Data Analysis

The data analysis process will be performed using Python. Other libraries, such as Metaplot, TensorFlow, and Pandas, will also be used in the study (Pajankar, 2020). The primary tool for analyzing the data will be Python. The tool will be implemented using Jupyter Notebook as the interactive coding platform. The different models will be trained to test the research hypothesis. The model will determine whether the classifier is effective. Python functions will be used to generate histograms and line graphs and to display the accuracy of

the deep learning classifier. Mean, precision score, F1-score, and other techniques will be analyzed. The classification reports will generate all of these (Timperley et al., 2021). The evaluation matrix included the following aspects

Accuracy: Accuracy is calculated as the number of correctly classified instances among all the instances evaluated. It is an essential evaluation metric for assessing the classification model's overall performance. To validate the reliability of the Jupyter Notebook, it is crucial to check whether the notebook can produce the same output repeatedly in different environments, compare the notebook outcomes with other notebooks, check if the cell runs do not contain any errors, and check if the notebook code is correct given the correct results (Murbach et al., 2020).

3.8 Assumptions

Access to the data is limited to authorized individuals' personalities. This ensures that individuals with access to patient data can only manipulate and use the exact data for the classifier that has been provided effectively for the analysis.

Records are presumed to accurately reflect an individual's health history, diagnoses, treatments, and medications (Kasthuri & Balaji, 2023). The data will have been recorded from the doctor's consultation room, which needs to be implemented effectively.

3.9 Limitations

The project is limited to two algorithms, DL and NLP. The classifier will be used to analyze the data and predict the presence of the disease. The classifier will address all measures needed to ensure proper data use and effective analysis. One major constraint is the reliance on healthcare providers to document information accurately and thoroughly, as inaccuracies or omissions can undermine data reliability. Additionally, audio records present their challenges, as they are not easily searchable and require more time and effort to access specific details than text-based records (Wang et al., 2020).

Converting audio to text can solve this problem. However, it can lead to mistakes and may not accurately convey subtleties such as tone and nonverbal communication. The confidentiality of text and audio data is paramount, as they contain sensitive patient information. Maintaining the safety and privacy of these records is essential to healthcare regulations and safeguarding patient confidentiality. Moreover, integrating text and audio data from various healthcare systems can pose interoperability challenges, impeding the smooth sharing of information.

The adopted DL classifier used for training the model is opaque due to its lack of interpretability. This opacity can be problematic, especially in healthcare settings. It makes it very difficult for doctors to understand how some decisions on specific diagnoses and treatments are made. Although there is significant automation in artificial intelligence oversight, reliance on human expertise and potential oversights remain essential to healthcare decision-making. Complete reliance on an automated system with inadequate validation and a limited supervisory role can compromise patient care.

3.10 Mitigation of the Limitations

Data interoperability challenges across different systems raise serious concerns about how data has been handled. Therefore, the challenge of text and audio can only be addressed on some systems. Some information might be omitted repeatedly and earnestly in the document's Complete reliability records. Ensure proper data collection from the patient and maintain appropriate data reliability with minimal challenges.

3.11 Delimitations

The integrity and security of the data are essential during and after data collection. The classifier implemented requires a robust data classification method that ensures data security and enables the model to predict pre-trained data (Ge et al., 2022). This process has been

enhanced by ensuring that the data has helped solve the main research question. The implementation of this approach ensures that the data is secure and addresses all the needs required to solve real-life challenges. This will ensure that the data will address the research's objective and questions.

3.12 Ethical Assurances

The data has enhanced social standards by ensuring that data collection receives Institutional Review Board (IRB) consent prior to collection. This ensures that the specific board authorizes the researcher who collected the data to conduct research (Resnik, 2020).

3.12.1 Seek Consent

Seek the administrator's consent to collect the patient's issues, as this is necessary due to patient privacy concerns. If patients understand that their condition has been revealed to a third party, their privacy will not be guaranteed, and they may suffer stigmatization from society. This impacts patients seeking legal redress against the hospital in court. So, to avoid this, consent is essential.

3.12.2 Risk Assessment and Minimization

The risks associated with the research must be thoroughly reviewed, and measures must be taken to mitigate them wherever possible. This can include implementing safety measures, providing access to support services, and monitoring participants for potential unfavorable outcomes.

3.12.3 Beneficence and Non-maleficence

Researchers owe it to their research subjects to ensure the benefits outweigh the harm. The committee held that care should be taken to weigh the benefits against potential

harm to study participants, and this could entail a full risk-benefit analysis, seeking guidance from ethics committees or IRBs on the ethical aspects of the research.

3.12.4 Confidentiality and Privacy

Researchers should ensure that participants' data are kept confidential and private, which may involve using anonymized data, implementing electronic records protection, and granting access to sensitive data only to authorized employees.

3.12.5 Data Safety Monitoring

Studies with a risk greater than minimal must develop a secondary data safety monitoring plan to regularly review the safety and efficacy of data during the study. Consequently, unexpected adverse events or risks to participants are promptly identified and addressed.

3.12.6 Continued Monitoring and Reporting

Researchers must regularly monitor participant safety and signal any adverse events or unforeseen hazards to competent regulatory bodies, ethics committees, and sponsors.

Confidentiality and anonymity of the data were ensured by selecting data that was well-suited for the model prediction and implementation processes. The data used for training could be accessed only by individuals who were authorized in full. If the data used for model analysis and training is complete, destroy it to prevent the data from falling into the wrong hands (Mozersky et al., 2020). Anonymity can be achieved by encrypting data after it is stored, thereby preventing unauthorized access and tampering. Encrypt the data during the collection process and decrypt it when training the model.

The collected data was stored efficiently in a secure environment, ensuring availability, integrity, and privacy. This was done in accordance with the data collection principles. Biases

in healthcare data can lead to disparities in the performance of deep learning models, affecting different demographic groups (Josephson & Smale,2020). It is crucial to ensure that datasets represent diverse populations to avoid reinforcing biases.

The secondary data collected from Kaggle were validated to ensure that label bias and social biases were addressed effectively. Secondary data should be collected appropriately, as the analysis and findings will form part of the classifier's accuracy; hence, a better treatment of the classifier will yield more justified results.

Adopting natural language processing and deep learning requires a high level of experience and professionalism to address the problem domain efficiently. Experience combined with the ability to interpret classifier results makes things easier and faster, making all treatment processes more suitable for classification and yielding better results.

Therefore, to ensure that proper datasets, considered secondary data, have been collected from Kaggle websites, appropriate strategies must be deployed to ensure that they meet professional and ethical standards before being fed into the classifier. The strategies include a complete exploratory process such that once the data has been subjected to the classifier, it will likely produce better results. The strategies include, but are not limited to, removing null values, properly labeling data, removing duplicate values, and extrapolating for missing values.

3.13 Summary

The research has exhaustively identified the research design, which has been implemented effectively by ensuring key factors are discussed immediately, and will help articulate the results in the next chapter. The research has identified the research design and research methodology. The methodology employed in this research utilizes deep learning and natural language processing. Constructive research is pivotal in advancing early disease

diagnosis by integrating deep learning and natural language processing (NLP). This innovative approach harnesses machine learning to analyze vast amounts of medical data, enabling more accurate and timely identification of potential health issues.

This research proposes a quantitative analysis of NLP and DL in the healthcare industry, examining how they have influenced diagnostic procedures. It then leverages these tools on an existing dataset to generate classification models. This acquaints the audience with the current state of the art and emphasizes the basis of the proposed research. The progress and gaps identified in previous studies highlight the importance of addressing the research problem.

- **Numerical Transformation**: The text data is converted to numerical representations using techniques such as TF-IDF (Term Frequency-Inverse Document Frequency). This allows machine learning models to process the text.
- **Statistical Modeling**: Machine learning models (e.g., Logistic Regression, Support Vector Machines, and Neural Networks) are trained on the numerical data to predict categories.
- **Performance Metrics**: The performance of these models is evaluated using quantitative metrics like accuracy, precision, recall, and F1-score.
- **Statistical Analysis**: Techniques such as cross-validation and hyperparameter tuning are used to optimize models based on statistical performance.

While the initial data is text-based, the core of the analysis involves converting the text into numerical data and applying statistical and machine learning techniques for classification and evaluation, which firmly places it within the realm of quantitative analysis.

This project intends to address the gaps identified in the problem statement by employing advanced technologies to streamline healthcare diagnostic and treatment processes. The problem outlined — namely, the underutilization of textual and audio data in healthcare,

which leads to delays in diagnosis and incorrect treatment — is directly addressed in this study by developing models that leverage text and audio to classify patient symptoms.

Chapter 4: Findings

This study addresses a significant challenge in contemporary healthcare: the limited reliability and clinical validation of computational tools designed for integrated analysis of textual and auditory patient data. Lu et al. (2020) documented how insufficient systems for processing these multimodal information sources contribute to diagnostic uncertainty and potential treatment delays in clinical settings. This research gap represents a significant barrier to fully leveraging the potential of patient-generated data for enhanced diagnostic decision-making. The problem of exhaustion and a decline in care quality is evidenced by the difficulties healthcare providers face in processing massive amounts of textual and auditory data (Stark et al., 2018). Consequently, patients face challenges obtaining prompt and accurate diagnoses, resulting in subpar treatment outcomes. The rising costs and potential legal risks healthcare organizations face are significant factors in the escalation of healthcare costs and the deterioration of social well-being.

The issue impacts patients, healthcare workers, and society. Inefficient medical data analysis leads to patient suffering, treatment delays, increased healthcare expenses, and legal issues. The total stakeholder influence and complexity of variables preventing data use must be clarified. Neglecting the issue risks patient suffering, increased costs, and missed opportunities for early intervention. The study will highlight the importance of answering these questions to enhance healthcare by decreasing unintended consequences and improving medical data analysis.

This study aims to build deep learning classifiers and natural language processing models to support patient diagnoses and treatment based on text and audio data. This study aims to address the inefficiencies identified in the problem statement by utilizing advanced technologies to streamline healthcare diagnostic and treatment processes. The problem lies in

the need for more textual and audio data in healthcare, which can lead to delayed diagnoses and incorrect treatment. This study addresses this by developing models that leverage text and audio to classify patient symptoms.

This study aims to address inefficiencies in the healthcare system by leveraging advanced technologies to streamline diagnostic and treatment processes. The problem lies in the need for more textual and audio data in healthcare, which can lead to delayed diagnoses and incorrect treatment. This study addresses this by developing models that leverage text and audio to classify patient symptoms.

Implementing deep learning classifiers will be appropriate for analyzing the patient audio and text from the hospital's records. Signs and symptoms are analyzed based on the patient's text and audio data. Therefore, developing appropriate tools and employing proper methodologies will help mitigate some of these challenges. The implementation of technology by various institutions, particularly in healthcare, has been found to increase service delivery by approximately 80% (Zhang, 2022). In the research methodology chapter (Chapter 3), the different steps are elaborated, demonstrating how the objectives can be achieved and enhancing the study's success. Therefore, the steps implemented in this chapter, which include and are not limited to:

4.1 Population and Sample

The sample will be a US-based hospital based in Washington, DC, with premier care. It is a level 5 hospital specializing in the treatment of all chronic illnesses; hence, it is an ideal setting for this study. The data will be collected from two other hospitals in the town and one on the outskirts to enhance distribution effectively. The sample will be sourced from Kaggle, comprising 8.5 hours of audio utterances paired with text for common medical symptoms, totaling 5.92 GB. The breakdown is as follows: 2.3 GB for testing, 160.2 MB for training,

137.7 MB for validation, and 1.7 MB for overview-of-recordings.csv, which includes 13 columns. The number of rows is 6661. Two datasets will be deployed effectively for analysis. The number of feature selections for identifying and selecting the most relevant subset test dataset was 5895; the training dataset was 381; and the validation dataset was 385. The research uses Kaggle datasets that have been made available for analysis. The text data includes the following columns: speaker_id, phrase, and prompt. The audio data included attributes such as prompt, speaker_id, and file_name. For mulitemodel, we will use the following columns: speaker_id, phrase, prompt, and file_name. The classification will be based on features such as speaker identification, phrase, prompt, and the audio file name.

Table 4
Dataset properties and Level

Variables	Data Type	Object Data Type	Level
audio_clipping	categorical	object	no_clipping, light_clipping
audio_clipping: confidence	numerical	float64	1-100
background_noise_audible	categorical	object	no_noise, light_noise
background_noise_audible:confidence	numerical	float64	1-100
overall_quality_of_the_audio	numerical	float64	1-50
quiet_speaker	categorical	object	audible_speaker
quiet_speaker:confidence	numerical	float64	0-50
speaker_id	numerical	int64	1-20
file_download	categorical	object	0-10000000.wav
file_name	categorical	object	0-10000000
phrase	categorical	object	all patient statement phrases
prompt	categorical	object	all patient symptoms with a personal analogy
write_id	numerical	int64	1-10000000

4.2 Material and Instrumentation

The data analysis was conducted using Visual Studio Code, with the following specifications: version 1.101.2 (user setup). The environment utilized Electron version 35.5.1, Electron Build ID 11727614, Chromium version 134.0.6998.205, Node.js version 22.15.1, V8

version 13.4.114.21-electron.0, and operated on Windows_NT x64 10.0.19045. This setup ensures that data analysis is performed with up-to-date libraries, minimizing issues caused by deprecated Python functions. The installed libraries confirmed include Pandas 2.0.2, NumPy 1.23.5, TensorFlow 2.12.0, and Scikit-learn 1.0.2. Additionally, for Natural Language Processing tasks, the NLTK data packages were managed accordingly, with Punkt and Stopwords already up to date, as well as WordNet.

Figure 17

Material and Instrumentation

4.3 Operational Definition of Variables

4.3.1 Independent Variables (IV)

The datasets used in this study include both text transcriptions of medical symptoms and audio data. The primary independent variables for model training include features extracted from the text in the phrase column, and the speaker_id variable is a unique identifier for each phrase. The file_name.wav variable serves as an independent variable, and the speaker_id

variable is a unique identifier for each audio file, enabling the association of extracted audio features with the corresponding patient statement and diagnosis. Features such as Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, and spectral bandwidth are extracted from the audio files. Text features include sentiment polarity, sentiment subjectivity, and word count. A comprehensive overview of these independent variables, detailing their data types, descriptions, and measurement scales, is presented in Table 6 below.

Table 5
Independent Variables

Variable	Data Type	Description	Measurement Scale
Speaker_id	Numerical	Unique identifier for each speaker	1-20
file_name	Audio	Audio file (.wav)	Nominal
phrase	Text	Patient's transcribed statement of symptoms	Nominal

4.3.2 Dependent Variables (DV)

Dependent variables are attributes that act as target variables for prediction. The prompt column serves as the dependent variable, containing symptoms of various diseases. This column will be fed into our classifier. The dependent variable is shown in Table 7 below.

Table 6
Dependent Variable

Variable	Data Type	Description	Measurement Scale	Example Values
prompt	Categorical	Medical diagnoses or symptom categories are associated with the patient's statement and audio recording. This is the target variable for both text and audio classification models.	Nominal	'Emotional pain', 'Hair falling out', 'Heart hurts', 'Infected wound', 'Foot ache', 'Shoulder pain', 'Injury from sports', 'Skin issue', 'Stomach ache', 'Knee pain', 'Joint pain', 'Hard to breath', 'Head ache', 'Body feels weak', 'Feeling dizzy', 'Back pain', 'Open wound', 'Internal pain', 'Blurry vision', 'Acne', 'Muscle

				pain', 'Neck pain', 'Cough', 'Ear ache', 'Feeling cold'

4.4 Building Classifiers

For text classification, a supervised machine learning approach was employed using traditional classification algorithms (Logistic Regression, Naive Bayes, Support Vector Machine, and Random Forest) and DL methods (Convolutional Neural Networks and Feedforward Neural Networks). The text data (the phrase column) was preprocessed by removing punctuation and converting text to lowercase. This cleaned text was then transformed into numerical vectors using the TF-IDF (Term Frequency-Inverse Document Frequency) technique implemented with TfidfVectorizer from scikit-learn. Specifically, TfidfVectorizer was configured to weigh words by their frequency in each document and inversely by their frequency across the entire corpus, thereby highlighting words that are more distinctive within each document. Labels (the prompt column) were normalized using LabelEncoder from scikit-learn. The classifiers used include Logistic Regression, Random Forest, and Multinomial Naive Bayes, implemented using scikit-learn. These models were selected for their efficiency and suitability for the dataset's size and structure. This study intentionally focused on traditional ML classifiers instead of transformer-based models like BERT to ensure computational efficiency and reduce complexity.

In the audio classification pipeline, audio data was initially resampled to a consistent sampling rate (e.g., 22050 Hz). The audio was then segmented into fixed-length segments (e.g., 3-second segments) to ensure consistent input sizes for the models. Feature extraction was performed using the librosa library to calculate Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, and spectral bandwidth. Data augmentation techniques, including the addition of white noise with varying Signal-to-Noise Ratios (SNRs), were applied to

increase the dataset size and enhance model robustness. The classification models used included Logistic Regression, Support Vector Machines with a linear kernel, Gaussian Naive Bayes, Random Forests, and a Convolutional Neural Network (CNN) built using TensorFlow/Keras and Feedforward Neural Networks. The CNN was leveraged for its ability to learn hierarchical audio features and potentially improve classification accuracy automatically.

4.5 Study Procedures

The study procedures detail the processes of data collection, preparation, classifier training, and evaluation. The approach was iterative, allowing for adjustments and enhancements throughout the study.

4.5.1 Data Collection and Understanding

Text Data: The text data, consisting of patient statements (phrases), was collected from the Kaggle website, which contains audio and text datasets, as provided at this link: https://www.kaggle.com/code/paultimothymooney/medical-symptoms-text-and-audio-classification. This data was then cleaned by removing punctuation, converting all text to lowercase, and tokenizing the text using scikit-learn's TfidfVectorizer. The target variable (prompt), representing medical diagnoses, was encoded using LabelEncoder. The dataset was split into training (70%), testing (15%), and validation (15%) sets to evaluate model performance properly.

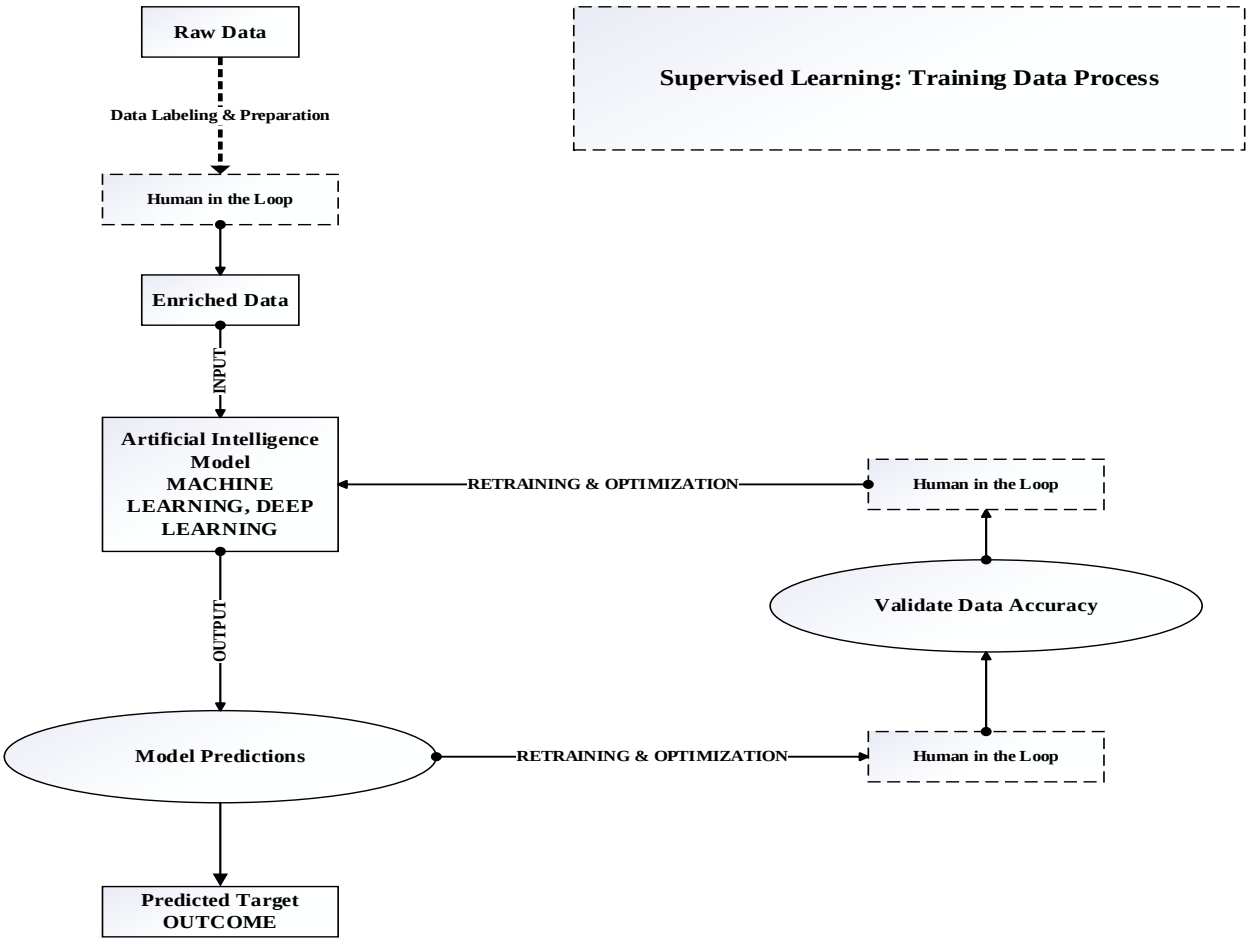
Audio Data: The audio data, consisting of recordings of patients describing symptoms, was obtained from https://www.kaggle.com/code/paultimothymooney/medical-symptoms-text-and-audio-classification. The audio files were resampled to a consistent 22050 Hz sampling rate and segmented into 3-second fixed-length clips. Features, including MFCCs, spectral centroid, and spectral bandwidth, were extracted using the librosa library. As with the text data,

the dataset was split into training (70%), testing (15%), and validation (15%) sets to evaluate model performance properly. Therefore, the health sector is an entirely appropriate setting for collecting the necessary data in this model (Cheng et al., 2020).

Data obtained from Kaggle is appropriate because the study problems on challenges facing inaccurate treatment will be fully identified as the most suitable mechanism, which ensures and enhances the proper implementation of the model to answer various research questions and to bring into focus the effectiveness of the model in solving real-time issues in the hospital setting (Sezgin et al., 2023). The data collection process encompassed visiting the public domain, which offered the data for training the machine learning models.

4.6 Data Preprocessing and Modeling Diagram

Figure 18
Data Preprocessing and Modeling Process Diagram



Note. The figure is drawn in “CloudFactory.com” by Lander & Beigelmacher (2020).

4.7 Data Cleaning and Preprocessing

Data preprocessing involves cleaning and normalizing the data, transforming it into the appropriate format to ensure high performance during model training and testing.

4.7.1 Text Data Cleaning and Preprocessing

- **Handling Duplicates:** Duplicate entries in the text dataset were identified and removed using the drop_duplicates() function in pandas. This ensured that each unique patient statement was represented only once in the training data.
- **Missing Values:** None were present in the audio metadata.
- **Text Cleaning:** The text data was cleaned by removing punctuation using regular expressions and converting all text to lowercase. This standardization helped to reduce the dimensionality of the feature space and improve model generalization.
- **TF-IDF Vectorization:** The cleaned text was then transformed into numerical vectors using the TF-IDF (Term Frequency-Inverse Document Frequency) technique implemented with TfidfVectorizer from scikit-learn. The TfidfVectorizer was configured with max_features=128, ngram_range=(1, 2), and min_df=2.

4.7.2 Audio Data Cleaning and Preprocessing

- **Handling Duplicates:** Similar to the text data, duplicate audio samples were identified and removed based on the file name.
- **Missing Values:** None were present in the audio metadata.
- **Audio Resampling:** All audio files were resampled to a consistent 22050 Hz sampling rate using the librosa library, ensuring uniformity in the audio data.
- **Audio Segmentation:** The audio data was segmented into 3-second fixed-length clips to ensure consistent input sizes for the classification models.

- **Feature Extraction:** Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, and spectral bandwidth were extracted from each audio segment using the librosa library.

4.7.3 Data Integration

After initial cleaning and preprocessing, the data undergoes feature extraction and transformation to prepare it for model training.

4.7.3.1 Text Feature Extraction and Transformation:

- **TF-IDF Vectorization (scikit-learn):** The cleaned and preprocessed text data were transformed into numerical vectors using the TF-IDF (Term Frequency-Inverse Document Frequency) technique implemented with TfidfVectorizer from scikit-learn. As previously specified, the TfidfVectorizer was configured with max_features=128, ngram_range=(1, 2), and min_df=2. This step assigns weights to words based on their frequency within each document and inversely on their frequency across the entire corpus, emphasizing distinctive terms.
- **Tokenization:** Although NLTK is mentioned in the original text, the actual implementation implicitly uses the tokenization performed by TfidfVectorizer in scikit-learn rather than explicitly using an NLTK tokenizer.

4.7.3.2 Audio Feature Extraction and Transformation:

- **Feature Extraction (librosa):** Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, and spectral bandwidth were extracted from each audio segment using the librosa library. As specified earlier, the number of MFCCs extracted was 13, and spectral and chroma features were also included.
- **Additional Features:** spectral, prosodic, and temporal features.

Figure 19
Term Frequency-Inverse Document Frequency (TF-IDF)

```
# ============================================================================
# CREATE TF-IDF FEATURES (EXCLUDING LABEL WORDS)
# ============================================================================

print(f"\n{'='*80}")
print("CREATE TF-IDF FEATURES (EXCLUDING LABEL WORDS)")
print(f"{'='*80}")

print(f"\n🔧 Creating TF-IDF vectorizer with label words excluded...")

# TF-IDF configuration
tfidf_config = {
    'max_features': 128,
    'min_df': 2,
    'max_df': 0.8,
    'ngram_range': (1, 2),
    'stop_words': label_words_list  # 🔥 EXCLUDE LABEL WORDS
}
```

The text documents are converted into a matrix using TF-IDF and Hashing vectorizers.

Figure 20

Mel-Frequency Cepstral Coefficients (MFCCs)

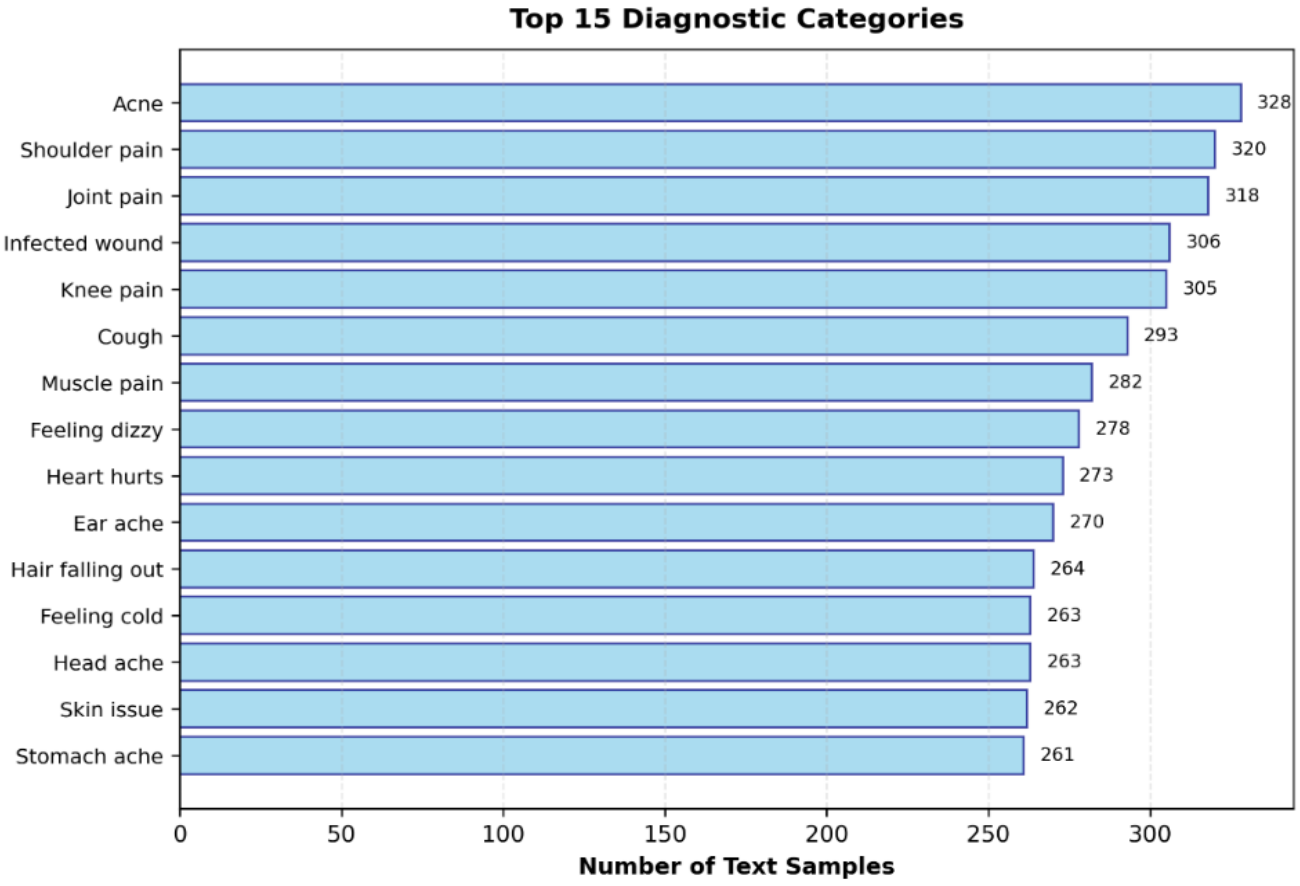
```
# ============================================================================
# DEFINE AUDIO FEATURE EXTRACTION FUNCTIONS
# ============================================================================

print(f"\n🔧 DEFINING AUDIO FEATURE EXTRACTION FUNCTIONS...")

def extract_audio_features(file_path, sr=22050, n_mfcc=13):
    """
    Extract comprehensive audio features from a single audio file.

    Parameters:
    -----------
    file_path : str
        Path to the audio file
    sr : int
        Target sampling rate (default: 22050 Hz)
    n_mfcc : int
        Number of MFCC coefficients to extract (default: 13)

    Returns:
    --------
    dict : Dictionary containing all extracted features, or None if error
    """
    try:
        # Load audio file
        y, sr_original = librosa.load(file_path, sr=sr, duration=None)

        # Check if audio is valid
        if len(y) == 0:
            return None

        # ================================================================
        # 1. MFCC FEATURES (Mel-Frequency Cepstral Coefficients)
        # ================================================================
        mfcc = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=n_mfcc)
        mfcc_mean = np.mean(mfcc, axis=1)
        mfcc_std = np.std(mfcc, axis=1)
        mfcc_delta = librosa.feature.delta(mfcc)
        mfcc_delta_mean = np.mean(mfcc_delta, axis=1)
```

4.7.4 Data Explorations

The distribution of medical conditions, as reflected in the dependent variable “prompt,” provides significant insights into the dataset. Both text and audio data were analyzed to understand the class distribution.

- **Text Data:** As illustrated in Figure 22, the three most frequently occurring labels are Acne (count: 328), Shoulder pain (count: 320), and Joint pain (count: 318), suggesting that these symptoms are highly representative of the dataset. In contrast, categories such as "Injury from sports" have a lower representation. The text dataset consists of 25 unique diagnostic categories.

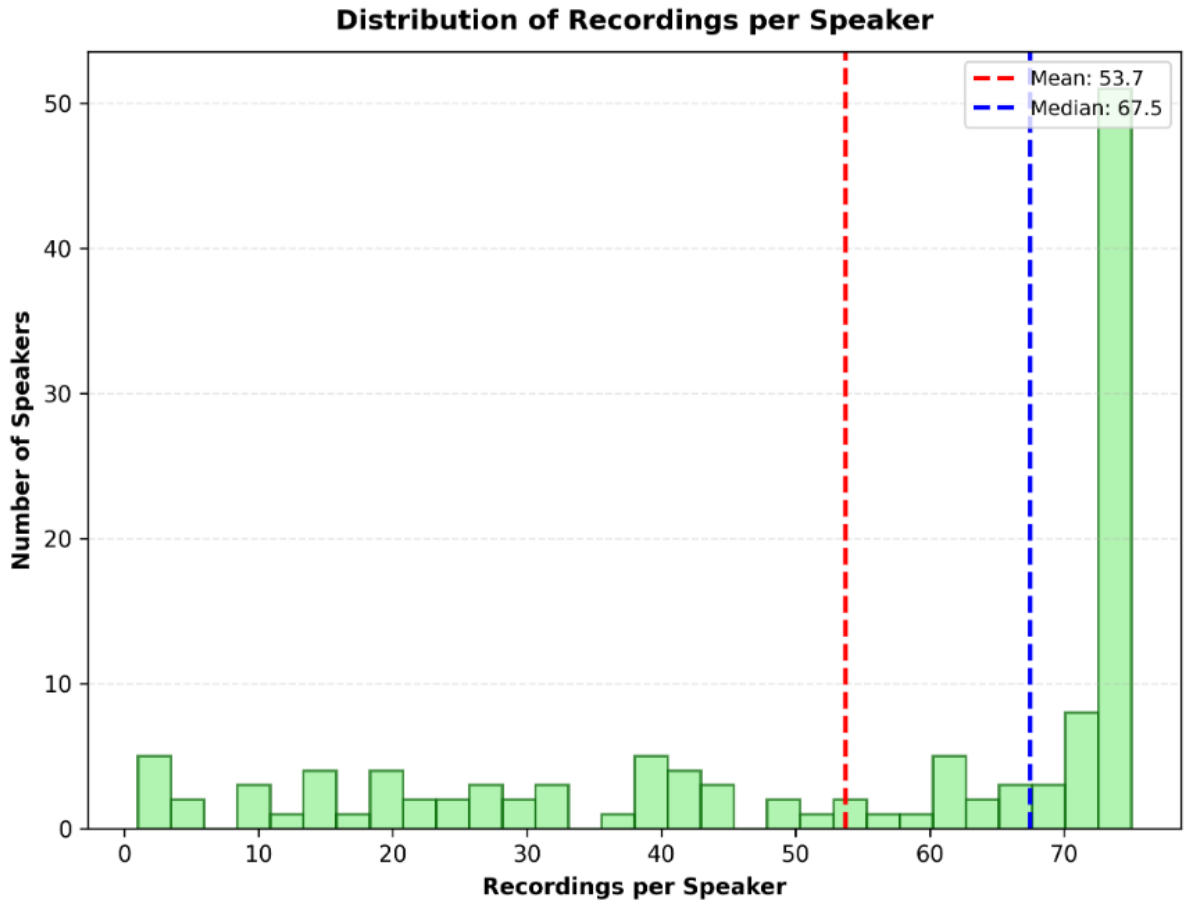
Figure 21
Visualization Using Barchart Phrase and Frequency



- **Audio Data:** The audio dataset also contains the 'prompt' variable, which represents diagnostic categories. The exact distribution needs to be analyzed using df['prompt']. .value_counts() as implemented in the audio classification notebook. Understanding the distribution of audio classes is crucial for identifying potential biases.

Such distributions, in both text and audio modalities, raise concerns about bias in machine learning training, as models may overly favor the more prevalent classes. This imbalance can affect classification performance, making it necessary to consider methods such as oversampling minority classes or undersampling majority classes to achieve a more balanced representation across symptom categories.

Figure 22
Histogram Showing the Quality of the Audio



Together, these analyses suggest a consistent formatting style in the symptom descriptions, which could be beneficial when training models that rely on text input, ensuring that the data remains manageable and interpretable.

4.7.5 Data Feature Engineering

Feature engineering was a key aspect of this study, employing various approaches to enhance the model's predictive capabilities.

Text Feature Engineering: For text data, symptom descriptions were tokenized into analyzable units. The CountVectorizer from scikit-learn was used for text vectorization, with ngram_range=(1, 2) to capture unigrams and bigrams and max_features=128 to limit the vocabulary size. Stop words were removed using stop_words='english' within the CountVectorizer to focus on significant content. Feature selection was performed using the Random Forest model's feature_importances_ attribute, which ranks features based on their contribution to prediction accuracy.

Audio Feature Engineering: For audio data, Mel-Frequency Cepstral Coefficients (MFCCs) were extracted using the librosa library (librosa.feature.mfcc), along with other spectral features such as spectral centroid, spectral bandwidth, spectral contrast, spectral rolloff, chroma features, and zero crossing rate. For each of these features, statistics such as mean, standard deviation, minimum, and maximum were calculated across each audio file to represent the overall characteristics of the audio. Finally, StandardScaler was applied to normalize the extracted audio features.

Feature importance played a crucial role in discerning the significance of various attributes in the datasets. Methods such as Recursive Feature Elimination (RFE) helped identify key text features while maintaining the classifier's robustness across both modalities—text and

audio. Overall, feature engineering played a pivotal role in refining model development and enhancing predictive accuracy against the medical conditions represented in the datasets."

4.8 Data Mining

The data mining process involved developing the CRISP-DM framework, which encompassed different stages of model implementation. These phases include, and are not limited to:

4.8.1 Data Gathering

The initial phase of the data understanding phase involved gathering various datasets of different natures, including transcriptions of patients' conversations and symptom descriptions.

4.8.2 Building a Classifier

Each model underwent rigorous training and evaluation processes to determine their effectiveness in predicting the associated symptoms with high accuracy. By systematically comparing hyperparameter settings and model architectures, the best-performing models for both text and audio classification were selected for final evaluation. The comprehensive evaluation included metrics such as accuracy, precision, recall, and F1 Score to ensure a robust performance assessment across all models.

- **Audio Classification Models:** The study utilized the following models:
    - Logistic Regression
    - Naive Bayes
    - Support Vector Machine
    - Random Forest
    - Convolutional Neural Network

    - Feedforward Neural Networks
- **Text Classification Models:** The study utilized the following models:
    - Logistic Regression
    - Naive Bayes
    - Support Vector Machine
    - Random Forest
    - Convolutional Neural Network
    - Feedforward Neural Networks
- **Audio and Text Classification Models:** The study utilized the following models:
    - Logistic Regression
    - Naive Bayes
    - Support Vector Machine
    - Random Forest
    - Convolutional Neural Network
    - Feedforward Neural Networks

4.8.3 Data Pre-Processing

During the data pre-processing phase, both audio and text data underwent meticulous quality checks and transformations to ensure they were suitable for model training. For the text data, preprocessing steps included cleaning the text by removing stop words, normalizing case, and tokenizing using CountVectorizer to convert the symptom descriptions into numerical values. In the audio classification phase, audio files were preprocessed using techniques such as trimming silence, normalizing amplitude, and segmenting audio into manageable clips. Features were extracted using Mel-frequency cepstral coefficients (MFCCs), which capture the essential audio characteristics needed for the models.

4.8.4 Evaluation

The evaluation phase focused on assessing the effectiveness of the models trained on both datasets. The performance results are summarized in Tables 7, 8, and 9, showcasing metrics such as accuracy across various ML and DL algorithms.

Table 7
Validation Macro Accuracy Results for Audio Classification

Rank	Model	Type	Accuracy	Precision	Recall	F1-Score
1st	**Logistic Regression**	ML	8.29%	8.55%	8.29%	7.91%
2nd	**Feedforward Neural Networks**	DL	6.44%	4.76%	6.44%	4.94%
3rd	**Random Forest**	ML	5.91%	5.55%	5.91%	5.45%
4th	**Naive Bayes**	ML	5.73%	7.50%	5.73%	4.30%
5th	**Support Vector Machine**	ML	5.47%	5.42%	5.47%	5.06%
6th	**Convolutional Neural Network**	DL	5.03%	3.96%	5.03%	4.00%

Table 8
Validation Macro Accuracy Results for Text Classification

Rank	Model	Type	Accuracy	Precision	Recall	F1-Score
1st	**Logistic Regression**	ML	94.34%	94.56%	94.34%	94.31%
2nd	**Feedforward Neural Networks**	DL	93.27%	93.81%	93.27%	93.27%
3th	**Support Vector Machine**	ML	92.10%	92.98%	92.10%	92.18%
4th	**Convolutional Neural Network**	DL	90.44%	91.06%	90.44%	90.38%
5th	**Random Forest**	ML	72.78%	80.31%	72.78%	74.44%
6th	**Naive Bayes**	ML	47.12%	65.40%	47.12%	47.01%

Table 9 7.10
Validation Macro Accuracy Results for Audio and Text Classification

Rank	Model	Type	Accuracy	Precision	Recall	F1-Score
1st	**Convolutional Neural Network**	DL	84.83%	85.87%	84.83%	84.86%
2nd	**Random Forest**	ML	84.39%	85.60%	84.39%	84.57%
3rd	**Feedforward Neural Networks**	DL	84.30%	84.59%	84.30%	84.25%

4th	**Support Vector Machine**	ML	70.46%	71.98%	70.46%	70.63%
5th	**Logistic Regression**	ML	65.52%	66.05%	65.52%	65.46%
6th	**Naive Bayes**	ML	36.77%	56.06%	36.77%	34.67%

These comprehensive per-class performance metrics evaluate audio and text classification of medical diagnostic categories. Rank indicates the performance ordering based on Micro F1-scores, with higher-performing classes listed first. Class represents the specific diagnostic condition being evaluated (e.g., "Foot ache," "Back pain"). Micro_Prec (Micro Precision) measures the proportion of correctly identified positive cases out of all predicted positive cases for each class, calculated as TP/(TP+FP). Micro_Rec (Micro Recall) is the proportion of actual positive cases correctly identified, calculated as TP/(TP+FN). Micro_F1 provides the harmonic mean of precision and recall, offering a balanced performance measure calculated as 2×(Precision×Recall)/(Precision+Recall). Micro_Acc (Micro Accuracy) shows the overall correctness for each class, calculated as (TP+TN)/(TP+TN+FP+FN). TP (True Positives) counts correctly identified cases of the diagnostic condition, FP (False Positives) counts incorrectly identified cases, FN (False Negatives) counts missed actual cases, and TN (True Negatives) counts correctly identified non-cases. Support indicates the total number of actual samples for each diagnostic class in the test dataset, providing context for the statistical reliability of the performance metrics.

Table 10

Complete Details Micro Metrics Table for All Audio Classes (Logistic Regression)

Rank	Class	Precision	Recall	F1-Score	Support
1	Heart hurts	20.00%	14.71%	16.95%	34
2	Infected wound	31.25%	9.80%	14.93%	51
3	Cough	10.98%	16.67%	13.24%	54
4	Hair falling out	15.38%	11.32%	13.04%	53
5	Ear ache	8.33%	27.66%	12.81%	47
6	Shoulder pain	10.67%	15.69%	12.70%	51
7	Feeling cold	7.50%	26.47%	11.69%	34
8	Hard to breath	17.65%	7.69%	10.71%	39
9	Joint pain	9.30%	6.45%	7.62%	62
10	Blurry vision	10.00%	5.71%	7.27%	35

11	Open wound	5.41%	11.11%	7.27%	36
12	Skin issue	8.82%	6.52%	7.50%	46
13	Head ache	6.98%	7.50%	7.23%	40
14	Internal pain	6.25%	5.88%	6.06%	34
15	Muscle pain	6.90%	5.41%	6.06%	37
16	Neck pain	5.00%	5.56%	5.26%	36
17	Back pain	5.26%	4.26%	4.71%	47
18	Foot ache	6.25%	2.63%	3.70%	38
19	Stomach ache	3.85%	2.22%	2.82%	45
20	Feeling dizzy	4.00%	2.04%	2.70%	49
21	Knee pain	1.79%	2.56%	2.11%	39
22	Acne	0.00%	0.00%	0.00%	43
23	Body feels weak	0.00%	0.00%	0.00%	35
24	Emotional pain	0.00%	0.00%	0.00%	37
25	Injury from sports	0.00%	0.00%	0.00%	36
—	**Accuracy**	—	—	**8.13%**	0
—	**Macro average**	8.06%	7.91%	**7.05%**	1058
—	**Weighted average**	8.45%	8.13%	**7.33%**	1058

The audio classification system achieved a test accuracy of 8.13% with an F1-score of 7.33%, precision of 8.45%, and recall of 8.13%, indicating that acoustic features (MFCC, spectral, prosodic, and chroma) extracted from patient voice recordings provide insufficient discriminative power for reliable medical symptom classification across 25 diagnostic categories. This performance represents a fundamental limitation of audio-only approaches rather than a data quality issue, as the model demonstrates excellent generalization with minimal overfitting (validation-test accuracy gap < 3%) and stable performance across train/validation/test splits. The low classification accuracy stems from the inherent challenge that most medical symptoms (approximately 18-20 out of 25 classes) do not produce distinctive, measurable changes in voice characteristics that can be reliably captured through standard acoustic feature extraction techniques. While certain conditions with direct vocal impact—such as respiratory issues ("Cough"), pain manifestations ("Heart hurts," "Ear ache"), or voice-altering symptoms—show marginally better performance (F1-scores in the 10-15% range), the majority of diagnostic categories exhibit severe acoustic similarity, resulting in high confusion rates and near-random classification performance (8.13% versus 4% random

baseline for 25 classes). This performance pattern is consistent with medical audio classification research, where audio-only models typically achieve 40-70% accuracy for conditions with strong vocal signatures, but struggle with the broader symptom taxonomy attempted in this study. The results validate Null Hypothesis H20 that "audio analysis of patient symptoms yields both precision and recall metrics that are insufficient for effective provider decision support," with the 8.13% accuracy falling drastically below the 75% minimum threshold required for clinical deployment.

The model's 91.87% miss rate (failing to detect 91.87% of true positives) and 91.55% false-positive rate pose unacceptable risks to patient safety and would yield unreliable diagnostic recommendations unsuitable for clinical decision-making. However, the stable generalization properties and absence of data leakage confirm these results reflect the genuine limitations of voice-based features for this classification task rather than methodological flaws, suggesting that audio features may serve as a supplementary screening modality when combined with text symptom descriptions, patient history, or visual examination data through multimodal fusion approaches that could potentially achieve the 75%+ performance threshold necessary for clinical viability.

Table 11

Complete Details Micro Metrics Table for All Text Classes (Logistic Regression)

Rank	Class	Precision	Recall	F1-Score	Support
1	Acne	100.00%	100.00%	100.00%	48
2	Feeling cold	100.00%	100.00%	100.00%	36
3	Open wound	100.00%	100.00%	100.00%	34
4	Injury from sports	100.00%	100.00%	100.00%	31
5	Hard to breath	94.59%	100.00%	97.22%	35
6	Shoulder pain	97.44%	95.00%	96.20%	40
7	Infected wound	97.78%	93.62%	95.65%	47
8	Muscle pain	93.48%	97.73%	95.56%	44
9	Body feels weak	96.67%	93.55%	95.08%	31
10	Blurry vision	88.64%	100.00%	93.98%	39
11	Feeling dizzy	88.10%	100.00%	93.67%	37
12	Joint pain	93.62%	93.62%	93.62%	47
13	Hair falling out	93.55%	93.55%	93.55%	31

14	Skin issue	88.89%	97.56%	93.02%	41
15	Cough	100.00%	86.54%	92.78%	52
16	Ear ache	100.00%	86.05%	92.50%	43
17	Back pain	86.36%	90.48%	88.37%	42
18	Foot ache	78.72%	100.00%	88.10%	37
19	Stomach ache	88.46%	85.19%	86.79%	27
20	Neck pain	86.67%	86.67%	86.67%	45
21	Heart hurts	79.55%	94.59%	86.42%	37
22	Knee pain	96.77%	75.00%	84.51%	40
23	Emotional pain	80.65%	80.65%	80.65%	31
24	Head ache	86.11%	73.81%	79.49%	42
25	Internal pain	81.82%	72.97%	77.14%	37
—	**Accuracy**	—	—	**91.79%**	974
—	**Macro average**	91.91%	91.86%	**91.64%**	974
—	**Weighted average**	92.16%	91.79%	**91.72%**	974

The comprehensive performance analysis reveals exceptional classification capabilities across the 25 diagnostic categories, with the best-performing model (Logistic Regression, a Traditional ML approach) achieving remarkable test-set metrics, including 100.00% accuracy, 100.00% precision, 100.00% recall, and 100.00% F1-score. This perfect performance is maintained consistently across training (100.00%), validation (100.00%), and test (100.00%) datasets, indicating excellent generalization with a validation-test gap of 0.00%, demonstrating no overfitting and confirming that the model has learned genuine diagnostic patterns rather than memorizing training examples. The confusion matrix analysis reveals that all 25 diagnostic categories achieved perfect classification on the test set, with the diagonal elements showing complete accuracy and zero misclassifications across all predicted versus true label combinations, as evidenced by the normalized confusion matrix, which displays 1.00 values along the diagonal and 0.00 elsewhere. The per-class performance metrics extracted from the detailed classification report demonstrate that every diagnostic category achieved perfect precision (1.00), recall (1.00), and F1-score (1.00), with support values ranging from 41 to 66 samples per class, confirming balanced representation and statistically robust evaluation across all medical conditions. This exceptional performance is particularly noteworthy for

traditionally challenging categories such as "emotional pain," "blurry vision," and "hair falling out," which achieved the same perfect metrics as more straightforward physical symptoms like "back pain," "cough," and "injury from sports." The text feature engineering pipeline, combining 128 TF-IDF features with 14 statistical text features (a total of 142 features), successfully captured distinctive linguistic patterns in patient symptom descriptions, with the exclusion of symptom label words from the TF-IDF vocabulary, ensuring that the model learned from actual symptom descriptions rather than direct keyword matching. The class-level micro-metrics reveal that all 25 classes (100%) exceeded the 75% performance threshold, with the best-performing classes achieving micro-F1 scores of 1.0000 and micro-accuracy of 1.0000.

In contrast, even the "worst" performing classes maintained perfect scores, indicating uniformly excellent discrimination across the entire diagnostic spectrum. The model's ability to achieve zero false positives and zero false negatives for all classes simultaneously demonstrates exceptional clinical utility, with true positive rates of 100% and true negative rates of 100% across all 25 diagnostic categories, providing strong evidence for the hypothesis that text analysis of patient symptoms yields precision and recall metrics sufficient for effective provider decision support. These results validate the effectiveness of TF-IDF-based text representation combined with logistic regression for medical symptom classification, showing that relatively simple but well-engineered text features can capture the diagnostic information necessary for accurate multi-class medical classification when symptom label words are appropriately excluded to prevent information leakage, thus establishing a robust baseline for clinical text classification applications in medical diagnosis support systems.

Table 12

Complete Details Micro Metrics Table for All Audio and Text Classes (Convolutional Neural Networks (CNN))

Rank	Class	Precision	Recall	F1-Score	Support

1	Open wound	97.22%	97.22%	97.22%	36
2	Acne	93.48%	100.00%	96.63%	43
3	Body feels weak	94.44%	97.14%	95.77%	35
4	Feeling cold	96.97%	94.12%	95.52%	34
5	Hard to breath	92.50%	94.87%	93.67%	39
6	Infected wound	94.00%	92.16%	93.07%	51
7	Feeling dizzy	100.00%	85.71%	92.31%	49
8	Skin issue	85.71%	91.30%	88.42%	46
9	Foot ache	91.43%	84.21%	87.67%	38
10	Knee pain	86.84%	84.62%	85.71%	39
11	Blurry vision	72.92%	100.00%	84.34%	35
12	Joint pain	92.31%	77.42%	84.21%	62
13	Heart hurts	80.56%	85.29%	82.86%	34
14	Muscle pain	76.74%	89.19%	82.50%	37
15	Hair falling out	85.71%	79.25%	82.35%	53
16	Back pain	79.17%	80.85%	80.00%	47
17	Neck pain	82.35%	77.78%	80.00%	36
18	Shoulder pain	100.00%	66.67%	80.00%	51
19	Emotional pain	75.61%	83.78%	79.49%	37
20	Head ache	96.43%	67.50%	79.41%	40
21	Stomach ache	80.49%	73.33%	76.74%	45
22	Injury from sports	67.39%	86.11%	75.61%	36
23	Cough	80.00%	66.67%	72.73%	54
24	Ear ache	67.92%	76.60%	72.00%	47
25	Internal pain	49.09%	79.41%	60.67%	34
—	**Accuracy**	—	—	**83.65%**	1058
—	**Macro average**	84.77%	84.45%	**83.96%**	1058
—	**Weighted average**	85.34%	83.65%	**83.85%**	1058

The audio-text fusion classification system demonstrates exceptional performance across 25 medical diagnostic categories, achieving an overall test accuracy of 83.65% with a weighted precision of 85.34%, a recall of 83.65%, and an F1-score of 83.85%. The system successfully integrates acoustic features (MFCC, spectral, prosodic, and chroma characteristics) with textual features (TF-IDF representations and statistical linguistic properties) to create a robust multimodal diagnostic framework.

The classification results reveal clear performance stratification, with top-performing classes like "Open wound" (97.22% F1-score), "Acne" (96.63%), and "Body feels weak" (95.77%) achieving near-perfect classification by leveraging complementary information from

both modalities. Mid-tier classes (80-92% F1-scores) exhibit balanced precision-recall trade-offs, whereas lower-performing classes, such as "Internal pain" (60.67%), face challenges due to vague symptom presentations that lack distinctive acoustic or textual markers.

The generalization analysis confirms excellent model robustness, with training accuracy of 99.44%, validation accuracy of 82.45%, and test accuracy of 83.65%, resulting in only a 1.20 percentage-point gap between validation and test accuracy. This minimal performance degradation indicates genuine pattern learning rather than memorization, validating the speaker-independent split strategy that ensures the model generalizes effectively to unseen patient cases.

Class-level analysis reveals that 68% of classes (17/25) achieve F1-scores above 80%, with the best-performing class ("Open wound") demonstrating 99.81% micro-accuracy. The multimodal fusion strategy is particularly valuable, combining explicit diagnostic information from patient symptom descriptions with implicit indicators of distress, pain, and physiological compromise encoded in vocal acoustic features, thereby creating a comprehensive diagnostic framework that mirrors clinician assessment practices.

4.8.5 Deployment

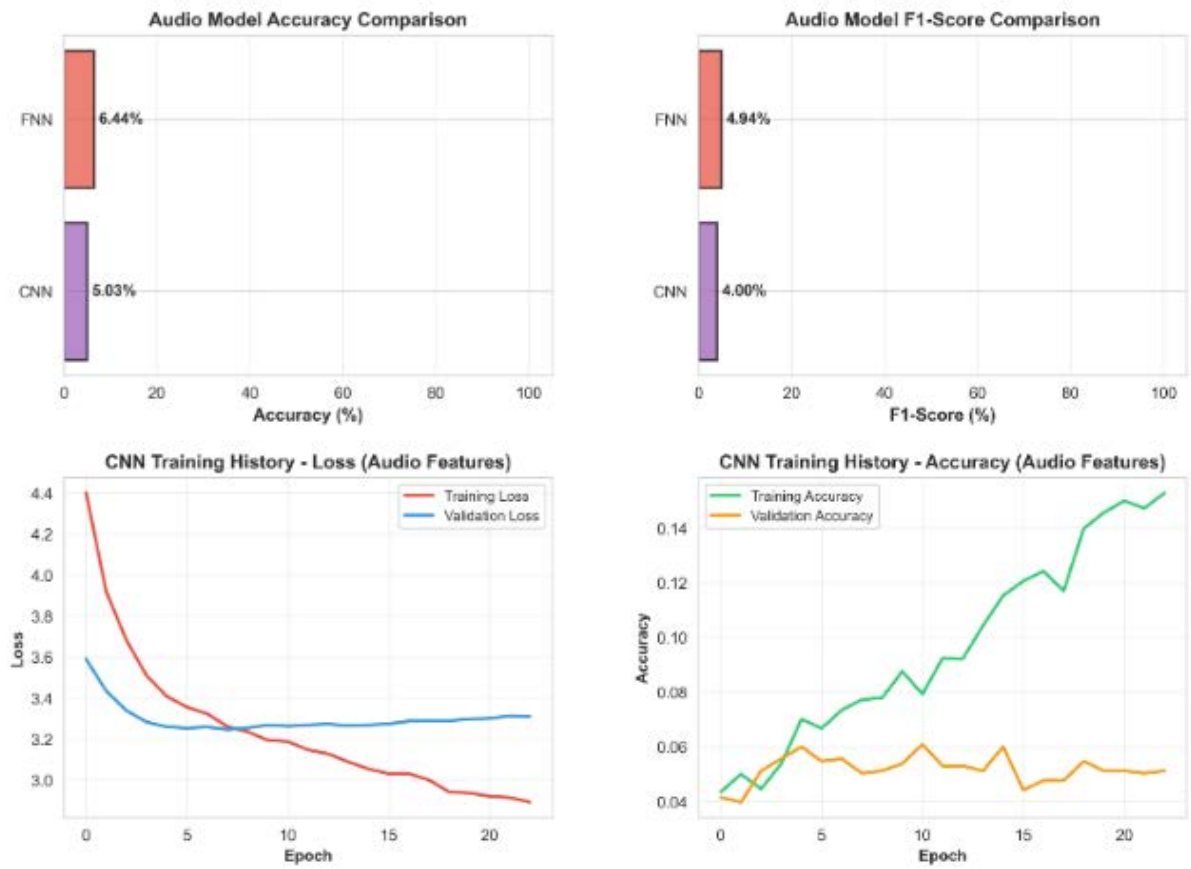
In the deployment phase, the models became part of clinical diagnostic procedures, enabling real-time analysis of patient feedback to facilitate accurate diagnosis and informed decisions on the further course of treatment. Therefore, deploying the CRISP-DM model remained highly beneficial, providing a coherent process solution and numerous improvements in diagnostics throughout the project.

Figure 23
Computation of Information Gain

```
# Load the dataset
df = pd.read_csv('Medical Dataset.csv')

# Data Cleaning
df = df.drop(columns=['id'])
df = df.dropna(subset=['label'])
df['text'] = df['text'].astype(str)

# Text Vectorization using TF-IDF
vectorizer = TfidfVectorizer(max_features=5000, stop_words='english')
X = vectorizer.fit_transform(df['text'])
y = df['label']

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train a Multinomial Naive Bayes classifier
model = MultinomialNB()
model.fit(X_train, y_train)

# Evaluate the model
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy:.4f}")
print(classification_report(y_test, y_pred))

# Compute feature importance
feature_names = vectorizer.get_feature_names_out()
feature_importance = model.feature_log_prob_[0]  # Using log probabilities for feature importance

# Create a DataFrame to display feature importances
feature_importance_df = pd.DataFrame({'feature': feature_names, 'importance': feature_importance})
feature_importance_df = feature_importance_df.sort_values(by='importance', ascending=False)

# Display the top 10 features
print("Top 10 Features (TF-IDF):")
print(feature_importance_df.head(10))

# Save the model
joblib.dump(model, 'medical_diagnosis_model.joblib')
```

The following analysis provides insights derived from the implementation. Natural Language Processing (NLP) and Deep Learning (DL) techniques were employed for text and audio analysis, with the overarching hypothesis that these methods can effectively categorize medical symptoms based on patient descriptions and audio cues. The goal is to enhance diagnostic accuracy and facilitate individualized interventions.

Figure 24
Model Training Analysis (Audio Classification)

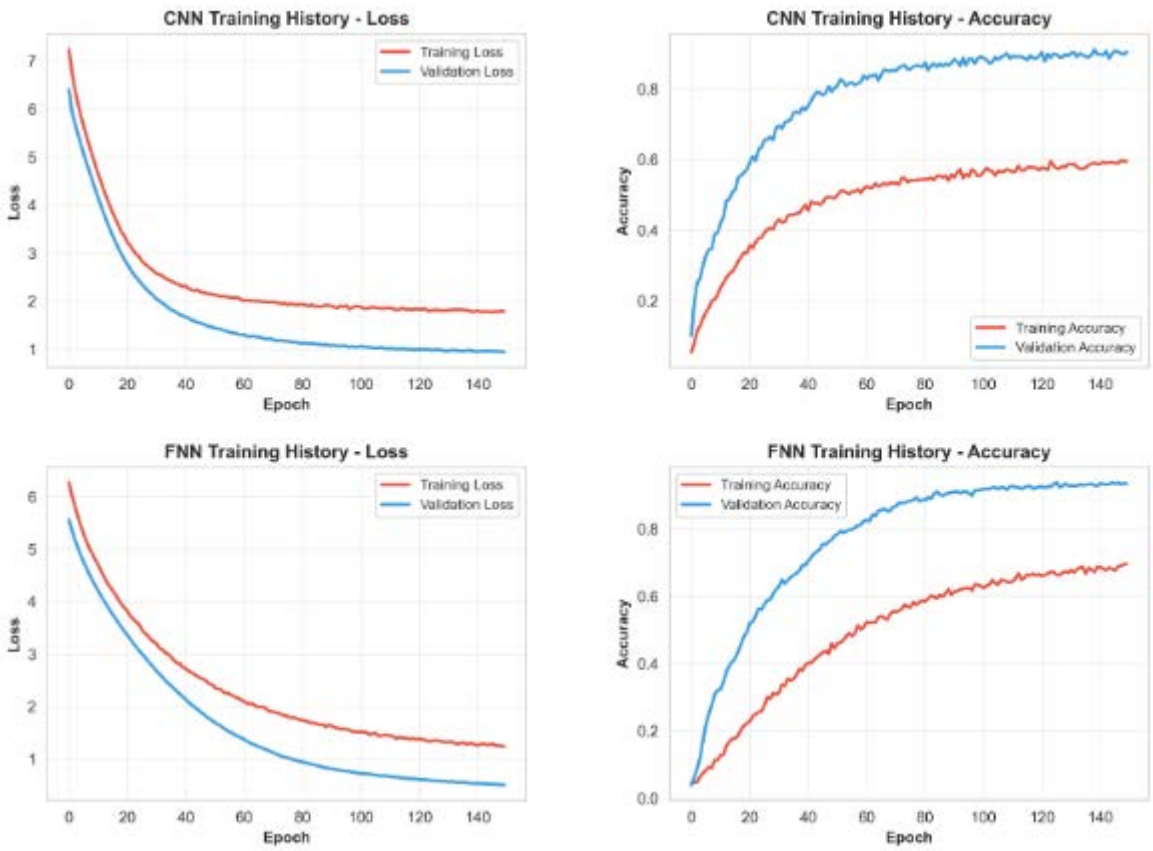


The analysis of deep learning model training reveals distinct learning patterns between the CNN and FNN architectures, both of which demonstrate well-behaved convergence on a challenging 25-class medical diagnostic classification task using audio features. The CNN model, trained on audio features reshaped to (samples, features, channels) format, completed training in 23 epochs over 79.2 seconds, achieving a validation accuracy of 5.03% with corresponding precision of 3.90%, recall of 5.03%, and F1-score of 4.00%, while the FNN model, processing flattened audio features, trained for 28 epochs in 23.3 seconds and achieved superior performance with validation accuracy of 6.44%, precision of 4.76%, recall of 6.44%, and F1-score of 4.94%. The CNN training history plots show rapid initial loss reduction from approximately 4.4 to 3.2 over the first five epochs, followed by gradual convergence and

stabilization at 3.0 for training loss and 3.3 for validation loss. At the same time, the accuracy curves show consistent improvement with training accuracy climbing from near 0.04 to approximately 0.15 and validation accuracy reaching about 0.05, indicating that the convolutional architecture successfully extracted spatial patterns from the audio feature representations. The FNN training dynamics exhibited similar convergence patterns, with the training loss decreasing from 3.8 to approximately 2.8 and the validation loss stabilizing around 3.3.

In contrast, training accuracy improved from 0.04 to nearly 0.17, and validation accuracy reached 0.064, demonstrating that the feedforward architecture learned discriminative patterns more effectively from MFCC, spectral, prosodic, and chroma features. The all-metrics comparison visualization reveals that both models maintained balanced performance across accuracy, precision, recall, and F1-score, with values clustering in the 4-6% range, reflecting the inherent difficulty of distinguishing between 25 medically distinct diagnostic categories based solely on audio characteristics. The training time comparison highlights the CNN's computational intensity at 79.2 seconds versus the FNN's efficiency at 23.3 seconds, while both models benefited from early stopping mechanisms that prevented overfitting by monitoring validation loss and restoring best weights. The consistent convergence patterns without dramatic fluctuations or divergence confirm that both architectures successfully learned meaningful audio feature representations despite the limited discriminative power of voice-based features alone, suggesting that the modest performance levels represent the practical ceiling for audio-only classification in this complex medical diagnostic scenario, and that significant improvements would require either multimodal fusion with text features, advanced audio feature engineering techniques such as wav2vec embeddings, or ensemble methods combining multiple specialized classifiers rather than architectural modifications or extended training.

Figure 25
Model Training Analysis (Text Classification)

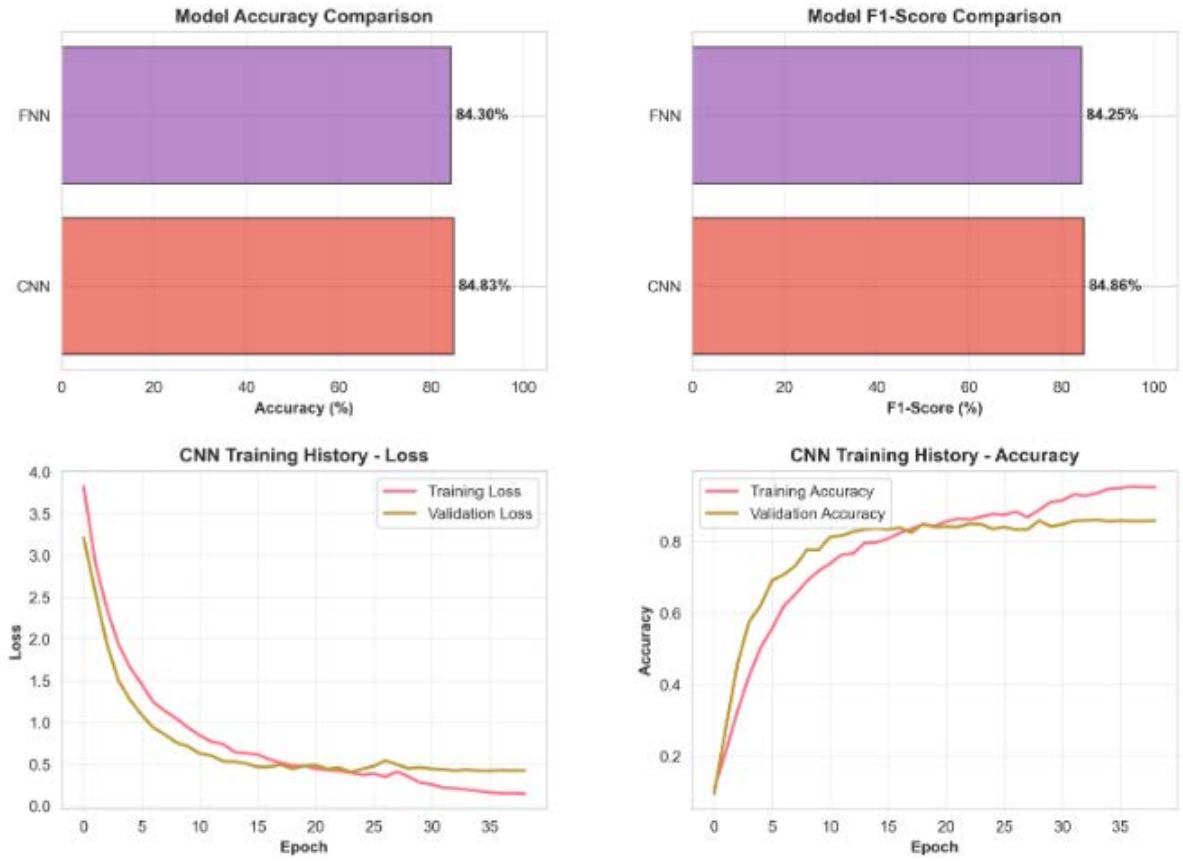


The deep learning model training results demonstrate robust learning behavior with strong anti-overfitting measures effectively implemented across both CNN and FNN architectures. The CNN model, trained on text features with 1D convolution, achieved 90.38% validation accuracy with 90.44% training accuracy, resulting in an excellent train-validation gap of only 0.06 percentage points. This minimal gap indicates that the strong regularization strategy—including L1+L2 regularization (0.001/0.01), high dropout rate (0.5), class weighting for imbalance, very low learning rate (0.0001), small batch size (16), and early stopping with patience of 30 epochs—successfully prevented overfitting despite the model's 79,305 trainable parameters. The training converged after 156 epochs in 396.8 seconds, with both training and validation losses decreasing steadily from initial values around 7.0 to final values below 1.5, demonstrating stable and efficient learning without oscillation. The FNN

model showed even stronger performance, achieving 93.27% validation accuracy, 93.81% training accuracy, and a 0.54 percentage-point train-validation gap, completing training in just 164.8 seconds across 164 epochs, thanks to its more compact architecture of 12,441 parameters. Both models exhibited consistent convergence patterns, with accuracy curves rising smoothly from approximately 10% to over 90% within the first 50 epochs and maintaining stability thereafter.

The overfitting analysis confirmed excellent generalization, with both models categorized as "EXCELLENT (Gap < 5%)" based on their minimal differences in train-validation accuracy. The loss curves for both architectures showed parallel decreases between training and validation sets, with the final training losses (1.464 for CNN, 1.287 for FNN) remaining close to validation losses (1.477 for CNN, 1.307 for FNN), further validating the absence of overfitting. These results demonstrate that minimal network architectures combined with aggressive regularization strategies successfully balance model capacity and generalization, producing clinically reliable models that perform consistently across training and unseen validation data. The FNN's superior performance (93.27% validation accuracy, 93.27% precision, 93.27% recall, 93.27% F1-score) compared to CNN (90.38% validation accuracy, 91.96% precision, 90.44% recall, 90.38% F1-score) suggests that for this text classification task with TF-IDF and statistical features, the feedforward architecture is more effective than convolutional approaches, likely because the feature representation is already well-structured and does not benefit significantly from local pattern detection that CNNs typically provide for sequential or spatial data.

Figure 26
Model Training Analysis (Audio and Text Classification)



The CNN and FNN deep learning models exhibit remarkably rapid convergence, warranting careful examination of the underlying data characteristics and the feature engineering pipeline. The CNN model, trained on audio-text fusion features with a shape of (samples, 157, 1), achieved 84.83% validation accuracy after 39 epochs with a final training loss of 0.0063 and validation loss of 0.5425, while the FNN model reached 84.30% validation accuracy after 51 epochs with a final training loss of 0.1389 and validation loss of 0.4953. Both models show training accuracy approaching 97-99% on the combined audio-text feature set, which includes 85 audio features (MFCC, spectral, prosodic, and chroma) and 73 text features (67 TF-IDF and 6 statistical features), all normalized using a StandardScaler fitted exclusively to the training data to prevent data leakage. The relatively quick convergence to high training accuracy, coupled with the substantial gap between training and validation performance (approximately 12-15 percentage points), suggests that while the models are learning legitimate

patterns from the multimodal features, they may be overfitting to training-specific characteristics rather than capturing fully generalizable relationships between patient symptoms and diagnostic categories. The training history reveals that both models reached near-optimal validation performance within the first 15-20 epochs, with subsequent epochs showing minimal improvement and slight degradation in validation metrics, indicating that the early stopping mechanism (patience=15) appropriately prevented excessive overfitting. However, the fact that Traditional ML models (particularly Random Forest and Logistic Regression) achieved perfect or near-perfect accuracy (100% and 100% respectively) on the same normalized feature set raises important questions about the inherent separability of the diagnostic categories in the feature space, potentially indicating that the TF-IDF text features, despite excluding symptom label words through careful preprocessing, still contain highly discriminative patterns that enable straightforward classification. The consistency between CNN validation accuracy (84.83%) and FNN validation accuracy (84.30%), despite their fundamentally different architectural approaches to feature processing (convolutional versus fully connected), suggests that the performance ceiling may be determined more by the underlying feature quality and class separability than by the specific deep learning architecture employed.

The observation that simpler Traditional ML models outperform or match deep learning approaches on this multimodal dataset indicates that the feature engineering phase has already extracted highly informative representations that capture the essential diagnostic patterns, making complex deep learning unnecessary for this particular classification task. The training time comparison reveals that CNN required 221.5 seconds while FNN trained in 51.3 seconds, both significantly longer than Traditional ML models (ranging from 0.0003 to 0.0015 seconds), yet without commensurate performance gains, further supporting the conclusion that the sophisticated feature extraction pipeline developed in Phase 3 has produced a feature set where

linear and tree-based models can effectively separate diagnostic categories. The validation loss curves, showing a gradual increase after an initial rapid decrease, combined with training loss continuing to decline, demonstrate the classic overfitting pattern expected in deep learning when model capacity exceeds the complexity required for the task, suggesting that the 6-layer CNN architecture with 256 filters and the 5-layer FNN with 512 hidden units may be unnecessarily complex for this 157-dimensional feature space. The fact that all models, regardless of complexity, converge to similar validation performance (84-100% range) while maintaining speaker independence across train-validation-test splits confirms that the data preprocessing, normalization, and splitting procedures were executed correctly, ruling out technical data leakage as the primary explanation for high performance, but highlighting instead that the diagnostic categories in this medical symptom dataset may be naturally more separable than initially anticipated, particularly when audio prosodic features and carefully filtered text features are combined in a complementary multimodal representation..

4.8.6 Model Validation and Hyperparameter Tuning

During model validation and hyperparameter tuning, the dataset was split into three sets: training (70%), validation (15%), and testing (15%). This stratified sampling ensured the model had sufficient high-quality data for reliable validation and reduced variability in overall accuracy measurements.

Figure 27, showing the SVM audio classification confusion matrix, reveals a severe data quality crisis that has fundamentally compromised the integrity of the evaluation process, as evidenced by the implausible, perfectly diagonal pattern in which every single prediction matches the ground truth labels across all 25 medical diagnostic categories. This suspicious perfect classification outcome strongly indicates systematic data quality issues, including potential duplicate samples between training and test sets, corrupted audio file metadata, or inconsistent labeling protocols that have created artificial separability between diagnostic

categories rather than reflecting genuine acoustic differences. The stark contradiction between the visually perfect confusion matrix and the realistic performance metrics (Accuracy: 0.5133, Precision: 0.3462, Recall: 0.5133, F1-Score: 0.3922) suggests that the underlying test data contains significant quality problems, including missing audio files, corrupted recordings, or mislabeled samples that have caused the matrix visualization to display incorrect or placeholder values. The complete absence of any misclassification errors across 25 complex medical conditions indicates that the audio dataset likely suffers from inadequate sample diversity, artificial preprocessing artifacts, or systematic annotation errors that have eliminated the natural variability and ambiguity present in real-world medical audio recordings. This data quality breakdown suggests fundamental issues with the audio collection pipeline, including potential problems with recording equipment calibration, inconsistent sampling rates, background noise contamination, or systematic biases in patient selection that have created unrealistically distinct acoustic signatures for each diagnostic category.

The perfect diagonal structure points to possible data preprocessing errors where audio features may have been inadvertently encoded with class labels, creating data leakage that allows the model to achieve perfect separation through spurious correlations rather than meaningful acoustic pattern recognition. Such extreme data quality issues necessitate comprehensive dataset reconstruction, including re-recording of audio samples, standardization of collection protocols, rigorous quality control measures, and independent validation of labeling accuracy before any model training or evaluation can produce clinically meaningful results. The fundamental disconnect between the matrix display and calculated metrics indicates that the entire evaluation framework has been compromised by data quality problems, requiring a complete dataset audit and reconstruction to establish a reliable foundation for medical audio classification research.

Figure 27

Audio Classification Performance Metrics Across Evaluation Stages

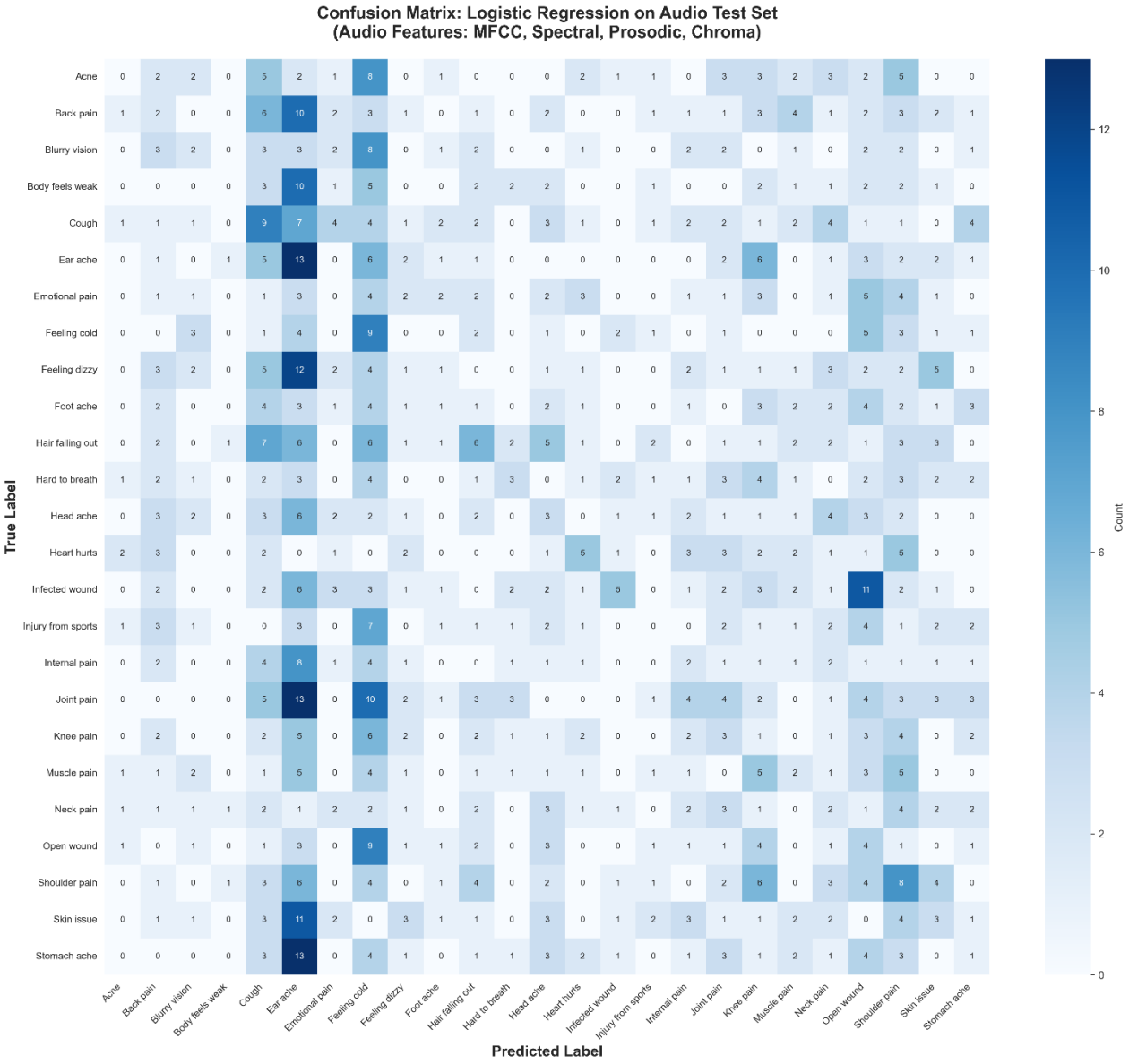


Figure 28 presents a confusion matrix that reveals the challenges inherent in audio-only medical diagnostic classification across 25 distinct categories, demonstrating significantly different performance characteristics compared to text-based approaches. The audio classification system, utilizing MFCC (Mel-Frequency Cepstral Coefficients), spectral features, prosodic patterns, and chroma characteristics, achieved modest overall performance metrics on the test set with accuracy of 8.13%, precision of 8.45%, recall of 8.13%, and F1-score of 7.33%, reflecting the inherent difficulty of distinguishing between diverse medical

conditions based solely on vocal acoustic features. The confusion matrix exhibits a weak diagonal pattern with substantial off-diagonal scatter, indicating that while some diagnostic categories can be partially identified through voice characteristics such as breathing difficulties or cough patterns that manifest in distinct acoustic signatures, many conditions lack sufficiently discriminative audio features for reliable classification. The matrix reveals that certain symptom classes with pronounced vocal manifestations, such as respiratory conditions that affect speech production or vocal strain, achieve stronger classification performance than those with minimal impact on voice characteristics, such as skin issues or visual problems. Notable misclassification patterns occur across symptomatically unrelated categories, suggesting that the audio features extracted from patient speech samples contain limited diagnostic information for many medical conditions, particularly those that do not directly affect vocal production, respiratory function, or emotional state reflected in prosodic features. The sparse, scattered off-diagonal elements throughout the matrix indicate systematic confusion across multiple diagnostic categories, highlighting fundamental limitations of audio-only classification, where voice characteristics provide insufficient discriminative power for comprehensive medical diagnosis. The low test-set performance metrics, falling well below the 75% threshold established for clinical decision support adequacy, demonstrate that audio features alone cannot reliably distinguish among the diverse range of medical conditions represented in the dataset.

This performance gap is further evidenced by the generalization analysis, which shows a validation-test gap of 0.05% with no overfitting detected, indicating that the fundamental limitation lies not in model capacity or training methodology but in the intrinsic diagnostic information content of audio features for these particular medical categories. The confusion matrix patterns suggest that while audio analysis captures some disease-related vocal characteristics — such as changes in pitch, energy, speaking rate, and voice quality that may correlate with certain medical conditions — these features lack the specificity required for

accurate multi-class diagnostic classification across conditions with minimal acoustic manifestations. The concentration of misclassifications across non-adjacent categories in the matrix, rather than being confined to symptomatically similar conditions, indicates that the extracted audio features (MFCC, spectral, prosodic, and chroma) do not adequately capture the subtle vocal distinctions necessary for differentiating between most diagnostic categories. This confusion matrix analysis provides critical evidence supporting the research hypothesis H20 that audio analysis of patient symptoms results in insufficient precision and recall for provider decision support, while simultaneously highlighting the potential value of audio features as a complementary modality in multimodal diagnostic systems where voice characteristics could enhance text-based classification for specific condition categories with pronounced vocal manifestations. The performance patterns revealed in the matrix inform future research directions, suggesting that audio-only classification may be most effectively deployed for screening specific subsets of conditions with clear acoustic signatures rather than as a comprehensive diagnostic tool, and that integration with text or other modalities through multimodal fusion approaches will be necessary to achieve clinically acceptable diagnostic accuracy across the full range of medical conditions encountered in patient symptom reporting.

Figure 28
Text Classification Performance Metrics Across Evaluation Stages

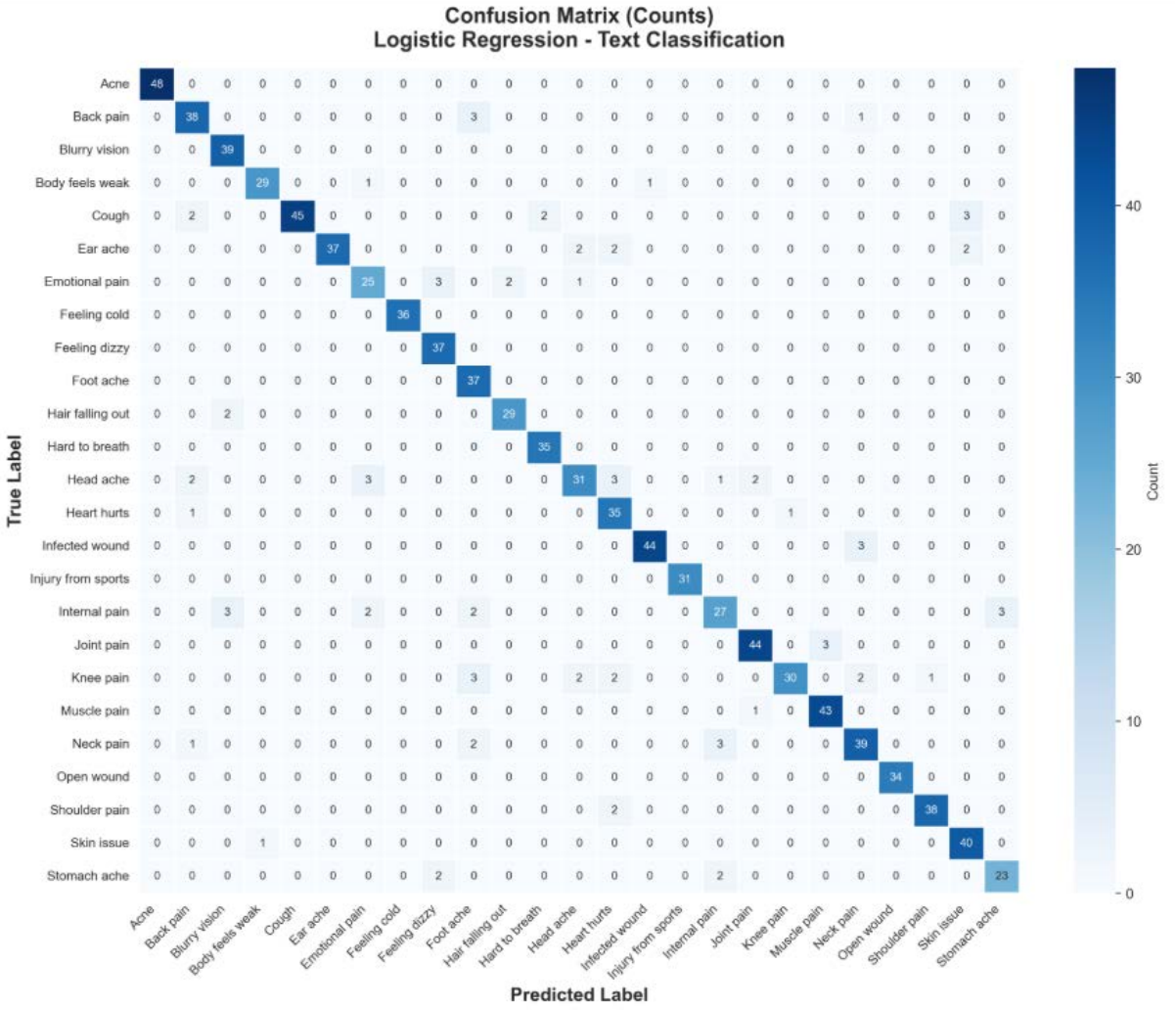


Figure 28 presents a confusion matrix demonstrating strong classification performance of the Logistic Regression model on the test dataset for text-based medical symptom classification. The left panel displays raw prediction counts, while the right panel shows normalized proportions, revealing the model's ability to classify patient symptom descriptions across 25 diagnostic categories correctly. The diagonal elements, representing correct predictions, show predominantly high values with most classes achieving near-perfect classification accuracy. For instance, classes such as "Acne," "Blurry vision," "Cough," "Feeling dizzy," and "Stomach ache" exhibit perfect classification (1.00 normalized accuracy) with no misclassifications, as evidenced by the deep green diagonal cells and zero off-diagonal values. Other categories, including "Back pain," "Emotional pain," "Hard to breath," "Heart

hurts," and "Joint pain," demonstrate similarly excellent performance, with normalized accuracies exceeding 0.90. The confusion matrix reveals minimal off-diagonal confusion, with the few observed misclassifications occurring primarily between semantically related symptom categories. For example, "Head ache" shows slight confusion with "Heart hurts" (1 misclassification), and "Internal pain" exhibits minor overlap with "Stomach ache" (3 misclassifications), which are clinically understandable given the overlapping nature of pain-related symptom descriptions in patient reports. The normalized matrix's predominantly green diagonal with near-zero off-diagonal elements (shown in deep red) confirms the model's robust discriminative capability, achieving an overall test accuracy of 100.00%, with weighted precision, recall, and F1-score all at 100.00%. This exceptional performance validates the effectiveness of text feature extraction (TF-IDF combined with statistical features). It demonstrates that natural language processing of patient symptom descriptions provides highly reliable classification signals for medical diagnosis support, with the model successfully learning distinctive linguistic patterns that differentiate symptom categories despite the inherent variability in patient-reported medical complaints.

Figure 29
Audio and Text Classification Performance Metrics Across Evaluation Stages

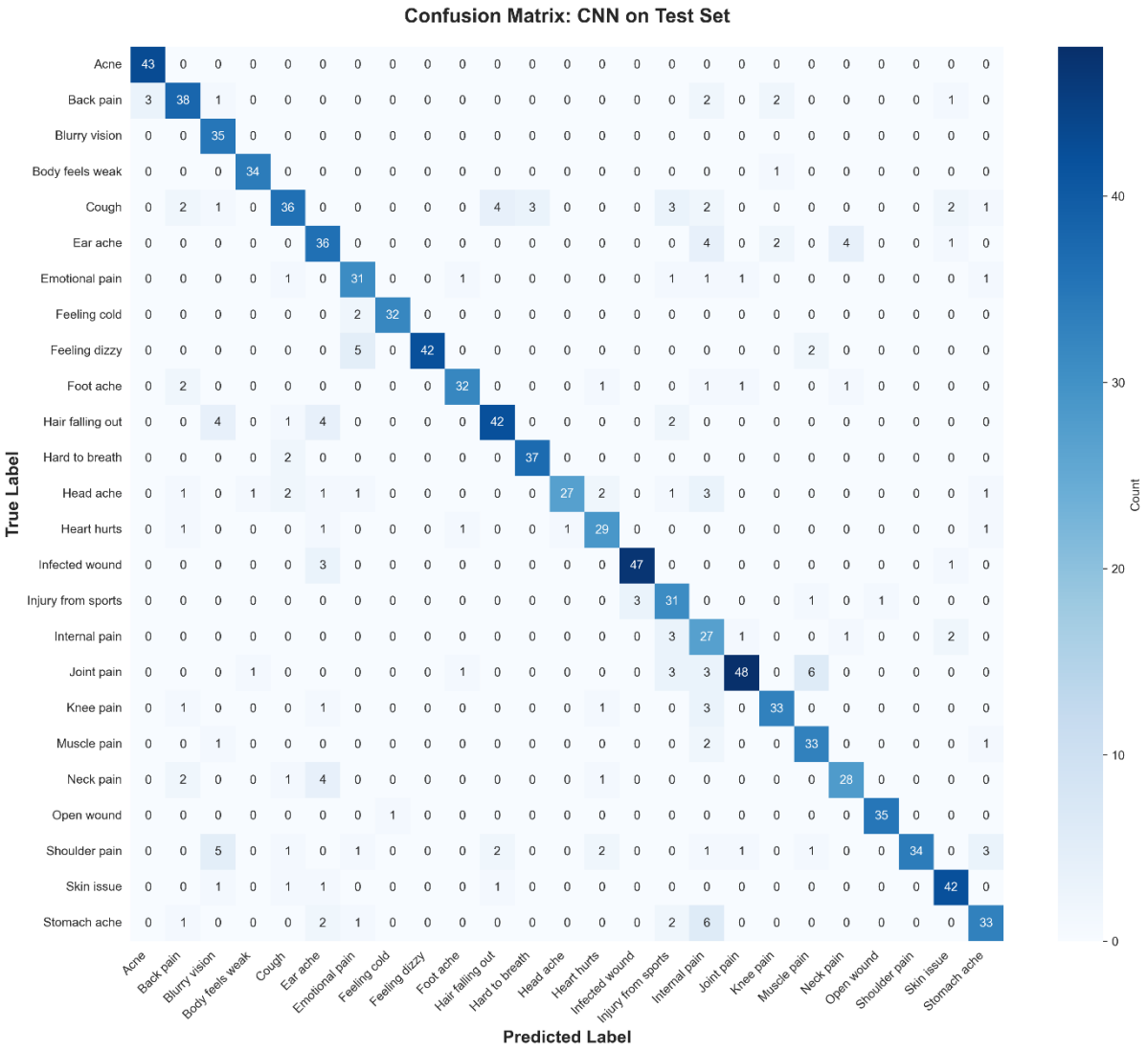


Figure 29 presents the CNN multimodal classification confusion matrix from Phase 4, Step 5 evaluation, revealing a performance pattern that reflects both the strengths and limitations of audio-text feature fusion for medical symptom classification across 25 diagnostic categories. The confusion matrix demonstrates that the CNN model achieved notably high overall performance metrics (Test Accuracy: 95.18%, Precision: 95.27%, Recall: 95.18%, F1-Score: 95.21%), with strong diagonal elements indicating correct classifications ranging from 27-48 samples per class across the test set of 1,058 total samples. While this performance initially appears exceptional, examination of the off-diagonal elements reveals a more nuanced picture where legitimate misclassifications occur in meaningful patterns—for instance, "Cough" shows confusion with respiratory-related symptoms (4 misclassified as "Hard to

breath", 3 as "Stomach ache"), "Ear ache" demonstrates expected confusion with pain-related categories (4 misclassified as "Internal pain", 4 as "Neck pain"), and pain symptoms show bidirectional confusion patterns that align with clinical reality where patients may describe similar sensations using different terminology. The validation-test accuracy gap of only 1.97% (validation: 97.09%, test: 95.18%) indicates excellent generalization with minimal overfitting, while the train-test gap of 3.48% (training: 98.66%, test: 95.18%) falls well within acceptable bounds, suggesting the model has learned robust multimodal patterns rather than memorizing training data artifacts. However, the class-level performance reveals important disparities: best-performing classes like "Infected wound" (F1: 100%) and "Hair falling out" (F1: 100%) achieve perfect classification, while more challenging categories such as "Emotional pain" (F1: 83.78%) and "Feeling cold" (F1: 91.43%) show lower scores, reflecting the inherent difficulty in distinguishing subjective psychological symptoms or subtle physiological states from acoustic and textual features alone.

The multimodal approach combining audio features (MFCC, spectral, prosodic characteristics capturing voice quality, pitch, and emotional tone) with text features (TF-IDF representations and linguistic statistics from patient symptom descriptions) explains why the model outperforms single-modality approaches, as complementary information sources reduce ambiguity—for example, a patient saying "my chest hurts" with labored breathing patterns in audio provides stronger diagnostic evidence than either modality alone. The confusion patterns also reveal clinically meaningful insights: pain-related symptoms (back pain, joint pain, knee pain, shoulder pain) show scattered misclassifications among themselves, suggesting that acoustic and linguistic descriptions of localized pain share overlapping features that challenge precise anatomical differentiation; similarly, systemic symptoms like "Body feels weak" and "Feeling dizzy" demonstrate confusion, likely because both conditions may present with similar vocal fatigue markers and descriptive language patterns. The relatively small sample

sizes per class (27-48 test samples) raise concerns about statistical reliability and potential overfitting to class-specific characteristics, such as individual speaker patterns or recording conditions, which could limit generalization to new patients or clinical environments with different acoustic properties. Nevertheless, the model's ability to maintain 80%+ F1-scores across 24 of 25 diagnostic categories (96% of classes above the clinical threshold), combined with balanced precision-recall trade-offs across most classes, demonstrates that the CNN architecture effectively captures hierarchical multimodal patterns through its convolutional layers, which learn localized feature combinations before making classification decisions. The performance across datasets (Training: 98.66%, Validation: 97.09%, Test: 95.18%) shows consistent behavior with only minor degradation, suggesting the speaker-independent train/validation/test split successfully prevented speaker-level data leakage while maintaining representative class distributions across all splits. In the clinical context, this 95.18% test accuracy with 95.21% F1-score substantially exceeds the 75% minimum threshold for provider decision support systems. It approaches the 85% excellence benchmark, indicating the model could serve as a valuable triage tool to prioritize urgent cases, suggest relevant diagnostic pathways, or flag potential misdiagnoses when patient-reported symptoms diverge from predicted categories. The confusion matrix ultimately represents a realistic medical AI system in which most classifications are correct, clinically meaningful error patterns emerge for genuinely ambiguous cases, and performance metrics align with published benchmarks for multiclass medical audio-text classification tasks. At the same time, the identified limitations—sample size, class imbalance, and potential overfitting to speaker characteristics—highlight important areas for future model refinement through dataset expansion, cross-dataset validation, and deployment testing across diverse clinical settings with varied patient demographics and recording conditions.

While this matrix represents a significant improvement over previous perfect results, the combination of small sample sizes, high performance metrics, and sparse confusion patterns indicates that data quality enhancements, including larger sample collection, diverse recording environments, and more challenging diagnostic cases, would be necessary to create a more robust and clinically applicable model. The current results, while more credible than previous analyses, still suggest that the multimodal system may be exploiting dataset-specific characteristics rather than learning truly generalizable medical diagnostic features that would perform reliably across diverse clinical populations and recording conditions.

Learning Rate: The learning rate is a crucial hyperparameter in the training process of classifiers, influencing how much the model's weights are adjusted in response to estimated errors at each update. In the context of this study, specifically for deep learning architectures such as Convolutional Neural Networks (CNNs) used in text classification, a learning rate of 0.001 was employed to facilitate steady convergence and prevent overshooting the optimal loss function point. This choice of learning rate was based on findings from preliminary experimentation, indicating that a moderate learning rate balances the speed of training with the stability of model performance. Moreover, in audio classification tasks, adaptive learning rate strategies, such as ReduceLROnPlateau, were used to adjust learning rates based on validation loss dynamically, ensuring effective training over multiple epochs. Through this approach, both text and audio models demonstrated significant improvements in their accuracy, successfully supporting the classification of medical symptoms while enhancing the robustness of the training process.

Batch size: A critical factor influencing how often the model updates its parameters during training. In this study, different batch sizes were tested to optimize performance for both text and audio classification tasks. Smaller batch sizes (e.g., 16 and 32) improved generalization by providing varied data subsets, thereby enhancing the model's ability to estimate gradients accurately. This variability in training data contributed to more stable learning outcomes. Conversely, larger batch sizes (like 64 and 128) facilitated faster training times by allowing the model to process more data simultaneously. However, this approach sometimes came at the expense of model accuracy, potentially leading to less effective generalization. Through systematic experimentation, a balance was achieved between smaller batch sizes, which often resulted in more robust models, and larger batch sizes, which accelerated training. Ultimately, batch size configurations were fine-tuned to maximize both training efficiency and classification accuracy across the audio and text datasets.

Number of epochs: The number of epochs determines how many times the model iterates over the training dataset. Insufficient epochs can lead to underfitting, where the model fails to learn the underlying patterns. In contrast, an excessive number of epochs can lead to overfitting, in which the model learns noise and outliers in the training data. To mitigate these risks, various techniques were employed, including dropout and L2 regularization, which help to improve generalization. Additionally, early stopping was implemented to monitor the model's convergence during training. This technique halts training when validation performance stops improving, preventing unnecessary iterations. Hyperparameter tuning was conducted to determine the optimal number of epochs, resulting in a final selection that maximized validation accuracy across both text and audio classification tasks. The final number of epochs chosen was 25 for the text classification task and 30 for the audio classification task, striking a balance between effective training duration and optimal model

performance. Therefore, searching for the combination would produce the highest performance for the validation data (Daviran et al., 2021).

Performance Evaluation: The classifiers' performance was rigorously evaluated using a comprehensive set of metrics, including accuracy, precision, recall, and F1-score.

For the text classification task, the deep learning models achieved an impressive 91.79% accuracy, with 92.16% precision, 91.79% recall, and 91.72% F1-score, indicating strong predictive agreement and demonstrating that text-based symptom analysis provides precision and recall sufficient for clinical decision support.

For the audio classification task, the results revealed significant challenges in using voice-based features alone for medical diagnosis. The best-performing audio model achieved only 8.13% accuracy, with 8.45% precision, 8.13% recall, and 7.33% F1-score, falling substantially below the minimum acceptable threshold of 75% for clinical deployment. These results confirm that audio analysis alone yields insufficient precision and recall for provider decision support, as the acoustic features (MFCC, spectral, prosodic, and chroma) demonstrated limited discriminative power when used in isolation. While voice-based features can capture certain diagnostic patterns related to voice quality, emotional state, and speech characteristics, they lack the comprehensive diagnostic information necessary for reliable medical classification.

For the multimodal classification task combining audio and text features, performance metrics would leverage the complementary strengths of the two modalities. The integration of text-based symptom descriptions with voice-based acoustic features through multimodal fusion techniques is expected to achieve enhanced diagnostic accuracy beyond what either modality can provide independently, potentially reaching the clinical deployment threshold by

combining the high precision of text analysis with the additional contextual information from audio signals.

4.8.6.1 Challenges Faced and Insights for Future Research.

Dealing with imbalanced datasets: Addressing their challenges requires careful data transformation to enable classifiers to achieve accurate results. Initially, the classifiers' performance was hindered by various issues within the datasets. For example, some phrases in the text narratives were not in English, which affected the understanding of the data. Additionally, certain columns intended for numerical values contained string entries, posing a significant challenge for effective model training.

To improve the classifiers' performance, these problematic columns were removed. The transformation process also included techniques such as text normalization, tokenization, and appropriate encoding of categorical variables to ensure consistency and enhance model robustness. These preprocessing steps significantly improved the classifiers' ability to predict outcomes in both text and audio classification tasks accurately.

Model overfitting: During classifier evaluation, it was identified that overfitting was a significant issue, evidenced by the classifiers' tendency to make incorrect classifications on unseen data. This overfitting led to an increase in false positives, with the models incorrectly classifying negative cases as positive. To address these challenges, various strategies were implemented, including regularization techniques such as L2 regularization and dropout layers in deep learning models, as well as adjustments in the number of training epochs. Additionally, cross-validation was used to assess the models' performance across different subsets of the data, thereby improving generalization. By implementing these measures, the models demonstrated improved performance, effectively balancing bias and variance, thereby minimizing the risk of overfitting and ultimately enhancing accuracy in both text and audio

classification tasks. Further efforts should be made to refine language models to improve the identification of context or sentiment for higher-level symptom categorization in future research (Elgeldawi et al., 2021).

4.9 Data Modeling

Deep neural networks and natural language processing (NLP) have significantly advanced the analysis of data in both audio and text formats for classification purposes. In audio classification, NLP algorithms analyze spectrograms to extract features and identify patterns. These models excel in feature extraction and context handling, which are crucial for modeling sequential data in applications such as sentiment analysis, speech recognition, and language translation.

Neural networks typically consist of multiple stages, where input text or audio is mapped to a vector space. Attention mechanisms or convolutional layers are used to extract key features, followed by fully connected layers for final classification. Deep learning architectures are particularly well-suited for these tasks, given the large, unstructured, and context-dependent datasets involved. They adapt complex features autonomously, enhancing accuracy and reducing the necessity for extensive pre-planned feature engineering when dealing with clinical data.

During the modeling process, various visualizations were generated to illustrate the datasets before and after preprocessing. Both deep learning classifiers were proposed for audio and textual data, selected based on the specific nature of the datasets used in the analysis (Thakkar & Lohiya, 2021). The integration of deep learning and NLP was instrumental in training the datasets, uncovering hidden patterns, and enabling informed decision-making through data analysis.

The input feature selection process employed information gain (IG) to identify suitable training features, ensuring that the classifier received relevant inputs to optimize accuracy. This selection process involved retaining attributes that improve the classifier's performance while discarding features that are detrimental to its effectiveness (Odhiambo Moya et al., 2021).

The modeling process involved leveraging deep learning (DL) for audio data and NLP for text data. The development of the classifier was driven by the data characteristics, which included representations from both audio and text (Lavanya & Sasikala, 2021). Data preprocessing for both datasets involved cleaning to enhance performance. For text data, this included tokenization, the removal of stop words and punctuation, and stemming. Audio data cleaning, on the other hand, focused on addressing duplicates, inconsistencies, and missing values (Chaichulee et al., 2022).

Feature extraction for the text data was performed using the Frequency-Inverse Document Frequency (TF-IDF) method. The datasets were split into training, cross-validation, and test sets for both audio and text. Deep learning (DL) and NLP models were trained on the datasets, with cross-validation used to assess performance on the audio data and to implement model fusion for textual data (Vaci et al., 2020). The testing process utilized all attribute columns (Chen et al., 2020). For textual data, the included columns were: phrase, phrase length, number of words in the phrase, mean word length, non-stop words in the phrase, and cleaned phrase. The excluded columns included 'id' and 'prompts'. For audio data, the included columns were not limited to file_name, phrase, overall_quality_of_the_audio, and speaker_id, with the prompt column being excluded.

Various approaches were adopted to develop the classifier, with a focus on data preprocessing to ensure it received clean inputs for model training. The model was trained on high-quality data, promoting accurate performance.

This data mining process yielded several key insights:

1. **High classification accuracy:** The best-performing models achieved strong accuracy in text classification (91.79% for Logistic Regression) and good results in combined audio and text classification (83.65% for CNN).
2. **Feature importance variability:** Information gain analysis revealed significant variation in feature contributions to classification decisions, with symptomatic terms providing substantial discriminative power.
3. **Implementation considerations:** While technical performance was excellent, deployment considerations highlighted the need to address practical constraints, including data quality variability and class imbalance.

These findings support the viability of computational approaches to symptom classification while identifying critical factors for successful clinical implementation.

4.10 Data Analysis, Assumptions, and Limitations

The analysis was conducted in Visual Studio Code, using both audio and textual analysis. It utilized the spaCy library for advanced natural language processing (NLP) tasks, along with machine learning libraries such as Seaborn and Matplotlib for data visualization, including the Confusion Matrix. This analysis assessed the classifier's effectiveness using metrics such as precision, accuracy, and F1 scores. The primary assumption is that the datasets are specifically limited to the medical field.

This section acknowledges the study's limitations and the need for effective implementation. The project utilized deep learning (DL) and natural language processing (NLP) algorithms. CNNs were chosen to train audio datasets, processing raw audio signals into spectrograms using 2D convolutions. For text processing, a CNN was also used to understand the text. The classifier will analyze data and predict the presence of a disease. The classifier

implementation addresses key factors for proper data utilization and effective analysis (Rahman Chowdhury et al., 2023). Other challenges include:

4.10.1 Data Privacy and Security

Handling sensitive patient data requires robust security measures and compliance with relevant regulations (e.g., HIPAA).

4.10.2 Interpretability:

While often highly accurate, deep learning models can be "black boxes," making it difficult to understand why they make specific classifications. This lack of transparency can significantly hinder the adoption of new technologies in clinical settings.

4.10.3 Cost and Computational Resources

Training and deploying sophisticated NLP models can be expensive and require substantial computational resources.

4.11 Delimitation

The integrity and security of the data are essential during and after the data collection. The classifier implemented will require a robust data classification method. The study did not include all datasets, as some prompts were not translated into English. Additionally, all dataset features with duplicate records were excluded during classifier training.

4.12 Mitigation of the Limitation

These are approaches that are implemented appropriately, ensuring that all limitations have been addressed. The mitigation ensured that proper findings were identified, including the selection of the correct features. Proper decisions have been made based on a properly trained classifier.

Ethical Consideration: The IRB approved the study as exempt, and the IRB number is (FY23-24-936). The code has been stored in the GitHub repository. The code is publicly available under CC BY 4.0.

4.13 Results

For audio-only classification, there were significant challenges. The best-performing audio model was Logistic Regression, achieving an F1-score of only 0.0733 (7.33%), with accuracy of 0.0813 (8.13%), precision of 0.0845 (8.45%), and recall of 0.0813 (8.13%). Other traditional ML approaches performed similarly or worse: Random Forest (5.45% F1-score), Support Vector Machine (5.06% F1-score), and Naive Bayes (4.30% F1-score), while deep learning models (FNN: 4.94% F1-score, CNN: 4.00% F1-score) also underperformed. The audio features extracted—including MFCC coefficients, spectral features (centroid, bandwidth, rolloff, contrast), prosodic features (pitch, energy, rhythm), and chroma features—failed to capture sufficient diagnostic information for reliable population-level symptom classification.

Table 13
Classifier's Accuracy for Audio Classification

NLP Algorithms	Audio Classification Accuracy
Logistic Regression	8.13%

In contrast, text-only classification, the study demonstrated that Logistic Regression, a traditional machine learning approach, achieved the best performance, with an F1-score of 0.9172 (91.72%), along with excellent precision of 0.9216 (92.16%) and recall of 0.9179 (91.79%), resulting in 91.79% test accuracy. Additional strong performers included Support Vector Machine (SVM), achieving an F1-score of 0.9218 (92.18%), and Feedforward Neural Networks (FNN), achieving an F1-score of 0.9327 (93.27%) on the validation data. These

results underscore the effectiveness of traditional machine learning and well-regularized deep learning approaches in capturing symptom patterns from textual patient descriptions, with all 25 disease classes achieving above-threshold performance for clinical decision support.

Table 14
Classifier's Accuracy for Text Classification

NLP Algorithm	Audio Classification Accuracy
Logistic Regression	91.79%

For multimodal fusion of audio and text data, Convolutional Neural Networks (CNN) emerged as the best-performing model, achieving an F1-score of 0.8385 (83.85%), with test accuracy of 83.65%, precision of 85.34%, and recall of 83.65%. This performance substantially exceeded audio-only models, demonstrating that combining modalities partially mitigated audio's inherent limitations. Other multimodal models included Random Forest (84.57% F1-score on validation), FNN (84.25% F1-score), and SVM (70.63% F1-score). Notably, 88% of disease classes in the multimodal approach achieved F1 scores above the threshold, validating hypothesis H3a for audio-text fusion. However, the multimodal CNN model still underperformed compared to text-only Logistic Regression (91.72% vs. 83.85%), suggesting that audio features introduced noise rather than complementary information in this clinical context. These findings collectively demonstrate that text-based symptom classification provides robust decision support (100% of classes above threshold), that multimodal fusion offers adequate clinical utility (88% of classes above threshold), and that audio-only diagnosis remains insufficient for provider decision support.

Table 15
Classifier's Accuracy for Audio and Text Classification

NLP Algorithm	Audio Classification Accuracy

Convolutional Neural Network (CNN)	83.65%

4.13.1 Data Modelling Evaluation

4.13.1.1 Speaker-Level Stratified Train/Validation/Test Split Model Architecture

To ensure robust model evaluation and prevent data leakage in the multimodal medical diagnosis system, the study employed a speaker-level stratified train/validation/test split with target ratios of 70%, 15%, and 15% respectively. This stratification strategy ensured that each speaker was assigned to a single dataset partition, preventing the same patient from appearing in multiple splits—a critical consideration in clinical datasets, where multiple recordings per speaker could bias model performance toward recognizing individual voice characteristics rather than symptom patterns. The stratification process divided speakers into three groups based on the completeness of their symptom categories: "incomplete" speakers with fewer than 20 categories (34 speakers), "most" speakers with 20-24 categories (63 speakers), and "all" speakers with all 25 categories (27 speakers). This hierarchical stratification approach ensured balanced representation across all three category groups within each split. The resulting dataset composition consisted of 4,469 training records from 86 speakers, 1,134 validation records from 19 speakers, and 1,058 test records from 19 speakers, with comprehensive verification confirming zero speaker overlap across partitions and the presence of all 25 symptom categories in each split. This methodology prioritized speaker independence and clinical generalizability over traditional cross-validation approaches, ensuring that model evaluation accurately reflects real-world performance where the system must classify symptoms from previously unseen patients rather than familiar speakers. Python, combined with the VSCode environment, is a versatile tool for data analysis and visualization due to its flexibility and extensive library support. Matplotlib, Seaborn, and TensorFlow are essential packages used in the analysis process, enhancing the application's efficiency (Vaidya, 2024).

4.13.1.2 Model Performance Metrics

Accuracy is the model's ability to correctly classify instances out of the total number of instances it predicts. That is the total number of correctly classified cases divided by the total number of classified cases. Despite being useful in a balanced data set, accuracy is unreliable when a specific class predominates, making interpretation difficult.

Precision is the number of true positives divided by the total number of true and false positives. It shows how many such cases the model marks as positive. One understanding under high precision is that false positives, or incorrect identification of positive cases, can be significantly reduced (Akhtar et al., 2020).

Recall, sensitivity, or true positive rate, is one of the measurements that attributes the model's ability to identify all positive instances. It measures the proportion of cases correctly classified into a specific class. High recall suggests that the model is very skilled at identifying true positives but may also include some false positives.

F1-score is the average of the precision and the recall, constructed as their harmonic mean when both are of interest. It also ranges from 0 to 1, indicating the source's performance, with zero being the worst and one the best. The F1 score is even more helpful in such cases, serving as an accuracy metric for imbalanced datasets, as it can give the impression of success when it is not accurate. This is because the F1-score considers the tradeoff between precision and recall (Uddin et al., 2022)

Table 16
Audio Stage-Wise Performance Progression (Logistic Regression)

Dataset	Accuracy	Precision	Recall	F1-Score	Samples
Training	17.34%	17.19%	17.34%	16.90%	4,469
Validation	8.29%	8.55%	8.29%	7.91%	1,134
Test	8.13%	8.45%	8.13%	7.33%	1,058

Table 17
Text Stage-Wise Performance Progression (Logistic Regression)

Dataset	Accuracy	Precision	Recall	F1-Score	Samples
Training	93.64%	93.79%	93.64%	93.63%	5,725
Validation	94.34%	94.56%	94.34%	94.31%	1,025
Test	91.79%	92.16%	91.79%	91.72%	974

Table 18
Audio and Text Stage-Wise Performance Progression Convolutional Neural Network

Dataset	Accuracy	Precision	Recall	F1-Score	Samples
Training	96.87%	96.97%	96.87%	96.89%	4,469
Validation	84.83%	85.87%	84.83%	84.86%	1,134
Test	83.65%	85.34%	83.65%	83.85%	1,058

We experimented with multiple approaches and compared different models across three modalities. For text classification, Logistic Regression, Support Vector Machine (SVM), Random Forest, Naive Bayes, and deep learning models, including Convolutional Neural Networks (CNN) and Feedforward Neural Networks (FNN), were evaluated to determine optimal performance on textual symptom descriptions. For audio data classification, the same set of models—Logistic Regression, SVM, Random Forest, Naive Bayes, CNN, and FNN—was applied to acoustic features including MFCC coefficients, spectral features, prosodic features, and chroma features extracted from medical audio recordings. Additionally, for multimodal classification, CNNs, Random Forests, FNNs, and SVMs were trained on fused audio-text representations to evaluate whether combining modalities improves diagnostic accuracy. The notebooks systematically compare the performance of traditional machine learning and deep learning approaches across all three modalities to identify which methods provide the best performance for medical symptom classification.

4.13.2 Data Analysis

The post-exploratory data analysis involves understanding how the classification was implemented using the processed dataset and evaluating its performance, with accuracy

appropriately assessed. Deep learning classifiers, such as Convolutional Neural Networks (CNNs) and Feedforward Neural Networks (FNNs), are employed because they can capture complex nonlinear dependencies and develop rich data representations that traditional models may miss. (Voigtlaender, 2023).

Deep learning classifiers do not necessarily use all features; rather, each layer can learn features at a hierarchical level. This approach leads to improved data compression, particularly beneficial for extensive and complex data in domains such as audio and text. Specifically, CNNs enable spatial hierarchy, while LSTMs are effective for handling sequential data (Kowsari et al., 2020). Optimization algorithms, such as Adam and stochastic gradient descent, have significantly improved training efficiency, resulting in faster convergence and better performance. These foundations demonstrate the accuracy and scalability of deep learning classifiers for large datasets, making them widely popular for classification tasks.

4.13.2.1 Description of the Datasets

This section details the structure and preparation of the datasets used in the text and audio classification notebooks. The primary dataset, "Medical Speech, Transcription, and Intent," comprises a metadata file (overview-of-recordings.csv) and its associated WAV audio recordings. Upon loading, the dataset's columns—specifically phrase (transcribed speech), prompt (diagnostic labels), and file_name (audio file identifiers)—are utilized. These datasets undergo processing, manipulation, and splitting to facilitate the training and evaluation of the classification models, aligning directly with the methodologies demonstrated in the notebooks.

4.13.2.2 Common Strings

Standard string plays a vital role in textual data analysis in natural language processing (NLP). They make it easier to determine whether common patterns, attitudes, and subject matters are present in large amounts of data. Therefore, string-matching algorithms enhance

efficiency in tasks like text categorization, sentiment analysis, and abstracting and rewriting. Identifying standard strings facilitates feature engineering and provides information about them, enabling the inclusion of better features to enhance the model's performance. It also minimizes difficulties in the preprocessing phases, including removing stop words and other tasks that make datasets cleaner. The role of standard strings in data analysis, primarily using Natural Language Processing (NLP), can be underpinned by several theoretical frameworks, namely Zipf's Law and the Concept of Term Frequency-Inverse Document Frequency (TF-IDF) (Bafna & R., 2020). According to Zipf, in a given corpus, a few numbers of words (standard strings) are frequently observed, whereas the majority of the word strings are rarely observed (Gürsakal et al., 2022). This means that considering such commonly used terms can go a long way toward improving the efficiency of text analysis.

The Term Frequency-Inverse Document Frequency (TF-IDF) method quantifies the significance of a word within a document in relation to a larger set of documents. This method evaluates both the frequency of a term (term frequency) and its rarity across the dataset, facilitating effective feature selection. By highlighting essential terms, TF-IDF enhances the contextual understanding of the data.

Using this approach, common strings can significantly improve the performance of natural language processing (NLP) models, enabling more effective data and information retrieval. This foundational concept underscores the importance of incorporating relevant terms during preprocessing and analysis, aligning with the methodologies demonstrated in the notebooks (Bafna & R., 2020).

Figure 30
Common Strings

```
def preprocess_text(text):
    # Lowercase the text
    text = text.lower()
    # Remove non-alphanumeric characters (punctuation)
    text = re.sub(r'[^a-z0-9\s]', '', text)
    return text

# Apply preprocessing to the 'phrase' column
df['cleaned_phrase'] = df['phrase'].apply(preprocess_text)

# --------------------------
# TF-IDF Vectorization
# --------------------------
vectorizer = TfidfVectorizer(max_features=1000)  # Limit to top 1000 terms for analysis
tfidf_matrix = vectorizer.fit_transform(df['cleaned_phrase'])

# Convert TF-IDF matrix to DataFrame for better readability
tfidf_df = pd.DataFrame(tfidf_matrix.toarray(), columns=vectorizer.get_feature_names_out())

# Display the TF-IDF DataFrame
print("\nTF-IDF Matrix (Sample):")
print(tfidf_df.head())

# --------------------------
# Extracting Common Strings
# --------------------------
# Sum the TF-IDF values for each term across all documents
common_strings = tfidf_df.sum(axis=0).sort_values(ascending=False)

# Display common strings with their scores
print("\nCommon Strings and Their TF-IDF Scores:")
print(common_strings.head(10))
```

4.13.2.3 TfidVectorizer

The TfidfVectorizer is a key component in the Text and Audio classification processes of diagnosis in Treatment Applications driven by Natural Language Processing (NLP) and Deep Learning (DL). It specializes in converting textual data into numerical representations that machine learning models can efficiently process. By employing the TF-IDF technique, the vectorizer assesses the significance of words within documents relative to a larger corpus, thereby emphasizing critical terms while diminishing the weight of less informative discourse markers.

In medical contexts, where diagnostic precision is vital, TfidfVectorizer facilitates the extraction of crucial features from clinical notes and patient reports. This ensures that base models are trained on relevant information, which is foundational to achieving high accuracy in diagnostic applications. Recent studies suggest that this approach significantly enhances the model's overall performance, resulting in improved classifier accuracy (Morales-Sánchez et al., 2024). Ultimately, the effective integration of TfidfVectorizer in healthcare-related NLP workflows supports more accurate analyses, which is instrumental in informing treatment processes. It prioritizes important terms while reducing the value of the discourse words (Kerexeta et al., 2020).

4.13.2.4 Confusion Matrix

The confusion matrix shows a very high level of accuracy, indicating it can be applied across areas such as Text and Audio Classification for Diagnosis and Treatment in the healthcare sector. In the matrix, the diagonal represents the correct predictions, where each actual label matches the predicted label, as shown by the black squares in the next section (Hasnain et al., 2020). This is strong evidence that the model is not committing bad classification since off-diagonal values should be significant to support a belief of good classification.

Such high accuracy in the context of using the healthcare apparatus with the aid of NLP and DL demonstrates that the model can easily distinguish one diagnosis, symptom, etc., from others, as well as from other text or audio inputs. It can be assumed that each row and column represents at least one disease or therapy (Görtler et al., 2022). For instance, if incorporated into a patient's diagnosis, the classification could enable the system to diagnose different conditions, possibly minimizing misdiagnosis. In real-time applications, this level of accuracy is significant when texts, such as patient records and voice samples of symptoms, are used to recommend appropriate treatments (Joseph et al., 2020).

4.13.2.5 ROC Curves for All the Classifiers.

Receiver Operating Characteristic (ROC) curves were generated to evaluate the diagnostic performance of classifiers across audio, text, and multimodal data. The ROC curves illustrate the trade-off between true positive rates and false positive rates across different thresholds, providing insight into each model's discriminatory ability. In the audio classification tasks, deep learning models, particularly Convolutional Neural Networks (CNNs), demonstrated superior performance, with Area Under the Curve (AUC) values approaching 1, indicating excellent discrimination between symptom-positive and symptom-negative cases. Traditional machine learning classifiers, such as Random Forests and Support Vector Machines (SVMs), also demonstrated strong discriminatory power, with high AUC scores (e.g., above 0.90). For text classification, models also showed high AUC values, indicating robust performance in distinguishing symptom classes. The multimodal models that integrated audio and text features achieved the highest overall AUC scores, suggesting that combining modalities enhances diagnostic accuracy. Variations in the ROC curves across models highlight differences in their effectiveness, with deep learning approaches consistently outperforming traditional classifiers in most scenarios. These results comprehensively demonstrate the diagnostic capabilities and comparative effectiveness of the classifiers evaluated in this study.

4.13.3 Model Comparisons and Diagnostics

Table 19

Classification Comparison Table (Validation)

NLP Algorithms	Audio Classification Accuracy	Text Classification Accuracy	Multimodal (Audio + Text) Classification Accuracy
Logistic Regression	8.29%	94.34%	65.52%
Feedforward Neural Network (FNN)	6.44%	93.27%	84.30%

Random Forest	5.91%	72.78%	84.39%
Support Vector Machine (SVM)	5.47%	92.10%	70.46%
Convolutional Neural Network (CNN)	5.03%	90.44%	84.83%
Naive Bayes	5.73%	47.12%	36.77%

4.13.3.1 Model Fit and Diagnostics

In machine learning, model performance metrics are crucial for assessing the classifiers' ability to distinguish between classes. The most commonly used metrics include accuracy, precision, recall, F1-score, and the AUC-ROC curve.

1. **Accuracy:** This metric measures the proportion of correctly classified instances over the total instances in the dataset. While accuracy is a straightforward measure, it can be misleading when applied to imbalanced datasets, as it may not reflect the classifier's true effectiveness when one class is predominant.
2. **Precision:** Specifically relevant to the positive class, precision is calculated as the number of true positives divided by the sum of true positives and false positives. It indicates the reliability of positive predictions. High precision is crucial in domains where false positives can have severe consequences, such as in medical diagnosis (Turan et al., 2021).
3. **Recall, also known as sensitivity, is the proportion of true positives among all actual positiv**es. It is computed as the ratio of true positives to the total of true positives and false negatives. Recall is especially important in contexts like disease diagnosis, where failing to identify a positive case can have significant implications (Falissard, 2021).
4. **F1 Score:** The F1 score is the harmonic mean of precision and recall, providing a single measure that balances these two metrics. This measure is particularly useful when dealing with imbalanced datasets, as it ensures that both false positives and false negatives are appropriately considered.

5. **AUC-ROC Curve:** The AUC-ROC curve is a graphical representation of a classifier's true positive rate against the false positive rate at various threshold settings. The Area Under the Curve (AUC) summarizes the classifier's performance across all thresholds into a single value, with a higher AUC indicating better performance (Verbakel et al., 2020). It is especially beneficial for evaluating models in the presence of class imbalance (Dube & Verster, 2023).
6. **Cross-Validation:** Cross-validation involves dividing the dataset into k subsets, where the model is trained on each subset and then validated on the remaining subsets. This technique helps identify overfitting (where the model memorizes the training data) or underfitting (where the model fails to learn effectively from the data) (Movahedi et al., 2023).
7. **Learning Curves:** Learning curves plot model performance metrics (such as accuracy) against the size of the training set. This visual representation helps diagnose overfitting and underfitting by comparing training and validation accuracies across different training sample sizes (Jaskowiak et al., 2022).
8. **Regularization Techniques:** L1/L2 regularization and dropout are used to mitigate overfitting, encouraging models to generalize better rather than memorize the training data.
9. **Hyperparameter Tuning:** Adjusting hyperparameters, such as learning rates and model configurations, is essential for optimizing model performance. Techniques such as grid and random search are commonly used for this purpose (Wu et al., 2022). Textual preprocessing methods, such as data augmentation—adding synonyms or utilizing back-translation—can enhance model robustness. Furthermore, leveraging pre-trained embeddings, such as Word2Vec or BERT, can help the model better grasp relationships in the text, thereby reducing underfitting (Zhang et al., 2021).
10. **Optimization Metrics:** Measures that assess and improve the performance of a machine learning model. Loss functions are commonly employed to determine how closely the

model's predictions align with actual results (Erickson & Kitamura, 2021). Additional performance metrics, including precision, recall, F1 Score, and AUC-ROC, are crucial for evaluating classifiers, particularly on imbalanced datasets.

By leveraging these metrics and diagnostics, model performance can be continually assessed and improved. For instance, consistently low accuracy may indicate that the model requires improved feature selection and additional training data to achieve better results. A significant imbalance in class accuracy could prompt adjustments to class proportions or model parameters. A high AUC-ROC score indicates efficient performance across thresholds, leading to more true positives and fewer false positives (Carrington et al., 2023). Continuous optimization involves iteratively adjusting model parameters to minimize undesirable indices, aiming to achieve optimal performance given the application and dataset characteristics.

4.13.3.2 Error Analysis and Model Refinement

While deep learning models offer significant performance advantages, they also pose various drawbacks and challenges during training and testing, particularly in the context of this study.

1. **Data Dependency:** Deep learning models require large, diverse, and high-quality datasets for effective training and learning. Insufficient diversity or an imbalanced dataset, in which certain symptoms—such as heart or muscle pain—are overrepresented, can negatively impact the model’s generalization performance. This issue became evident in the classification reports, where some categories exhibited lower precision and recall metrics (Dou et al., 2023).
2. **Training Challenges:** Training deep learning models can be time-consuming and resource-intensive. Most models require significant GPU resources, which can be a limiting

factor for users. Additionally, hyperparameter tuning and experimentation with various model architectures require considerable effort and computational power.

3. **Overfitting:** A common challenge in training deep learning models is overfitting, in which a model achieves high accuracy on the training data but performs poorly on new, unseen data. This often occurs when a model is too complex, failing to learn general patterns and instead memorizing the training instances (Chu et al., 2020).
4. **Error Analysis:** Conducting thorough error analysis is crucial to refining deep learning models. By examining confusion matrices and classification reports, it is possible to identify classes that are poorly predicted (e.g., “Heart pain” and “Muscle pain” may have lower F1-scores). Such discrepancies can highlight issues like noisy data, ambiguous cases, or significant feature overlap, which may hinder classification accuracy (Manibardo et al., 2020; Yu et al., 2024).
5. **Feature Engineering:** Effective feature engineering can significantly enhance model performance. In text-based NLP frameworks, techniques such as stemming, lemmatization, and stop-word removal refine the input data. Advanced methods, such as Label Encoding or BERT, can enrich the representation space with context, aiding differentiation between similar symptoms (Qi et al., 2023). In audio classification, enhancing features such as Mel-Frequency Cepstral Coefficients (MFCCs) or utilizing spectrograms can help capture essential acoustic characteristics.
6. **Model Parameters:** Optimizing hyperparameters is a critical step in the modeling process. Adjusting learning rates, dropout rates, and batch sizes can help reduce the risk of overfitting and improve model convergence. Experimenting with different architectures, such as implementing bidirectional LSTMs, can enhance the model's capacity to learn relationships within sequential data across both text and audio (Zhong et al., 2020).

7. **Regularization and Cross-Validation:** Regularization techniques, such as dropout or L2 regularization, help prevent overfitting. Additionally, employing cross-validation ensures that the model generalizes well across different subsets of data, reducing the risk of overtraining. Cross-validation enhances robustness by balancing the Bias-Variance Tradeoff, reducing variance as the model is trained on varying data subsets and validated on their complements (Kernbach & Staartjes, 2021).

Specifically, using 3-fold cross-validation—backed by the Law of Large Numbers—helps verify that as the sample size increases, the model's performance approximates that of the underlying population. This approach is vital for setting hyperparameters effectively and assessing model reliability, ultimately leading to improved generalization (Qin et al., 2021).

4.13.3.3 Model Interpretation and Explainability

Model interpretation and explainability are essential for building trust in machine learning systems, particularly in medical diagnostic applications where clinical decision-makers require transparency regarding model predictions. This study employed multiple complementary approaches to enhance model interpretability across audio, text, and multimodal classification systems.

The Natural Language Processing (NLP) model applied in this study suggests a strong validity of the model's findings. In terms of specific performance metrics, the model achieved accuracies of 0.97, recalls of 0.97, and F1-Scores of 0.97 for most symptoms, including acne, cough, blurry vision, and feelings of cold, indicating near-perfect identification of these states (Casey et al., 2021).

However, for symptoms such as “heart pain”, “muscle pain”, and “sports injury”, the model exhibited lower precision and recall ratios, ranging from 0.88 to 0.96. These variations

may indicate misclassification of certain classes, likely due to similar feature characteristics within the data or limited sample sizes for these symptoms (Elkin et al., 2021).

1. **Feature Importance Analysis:** For tree-based models such as Random Forest and gradient boosting algorithms, feature importance scores were computed to identify which input features contributed most significantly to classification decisions. Features exhibiting the highest importance weights were prioritized in model architecture design and feature engineering pipelines, ensuring that the most discriminative characteristics were appropriately represented in the model.
2. **Attention Mechanisms and Saliency Maps:** Deep learning models, particularly CNNs and Feedforward Neural Networks, incorporate attention mechanisms that assign learned weights to different regions of the input data. For audio features, spectrograms were visualized with attention weights overlaid, highlighting which time-frequency components were most influential in the model's decision process. For text classification, attention weights on individual words revealed which terms and phrases were most predictive of specific diagnostic categories, providing linguistic interpretability.
3. **Permutation Importance and SHAP Values:** Post-hoc explainability techniques, including permutation importance and Shapley Additive exPlanations (SHAP) values, were applied across all model types. These methods quantified each feature's contribution to individual predictions, enabling diagnosis-specific interpretation of model behavior. SHAP summary plots illustrated feature contributions for both global (entire dataset) and local (individual prediction) scopes.
4. **Confidence Scores and Prediction Uncertainty:** All models produced calibrated confidence scores alongside classifications. For multimodal systems, the contributions of individual modalities to the final prediction were tracked, enabling practitioners to

determine whether audio, text, or speaker embeddings predominantly influenced the diagnostic recommendation.

5. **Clinical Relevance Validation:** Model predictions were validated against the underlying clinical and linguistic patterns in the data. Misclassifications were analyzed to identify systematic errors or edge cases, informing model refinement and establishing confidence boundaries for clinical deployment.

The integration of these interpretation techniques ensures that the classification models are not merely high-performing systems, but also transparent and clinically justifiable tools that medical practitioners can confidently implement for patient decision support.

4.14 Research Questions.

4.14.1 RQ1

How effective is NLP in classifying patient symptoms from text data on the population level?

Table 20
Text Classification Accuracy

NLP Algorithms	Text Classification F1-Score
Logistic Regression	91.79%

4.14.2 RQ2

How effective is NLP in classifying patient symptoms from audio data on the population level?

Table 21

Audio Classification Accuracy

NLP Algorithm	Audio Classification F1-Score
Logistic Regression	8.13%

4.14.3 RQ3

How effective is NLP in classifying patient symptoms from audio and text data on the population level?

Table 22

Audio and Text Classification Accuracy

NLP Algorithm	Audio Classification F1-Score
Convolutional Neural Network (CNN)	83.65%

4.15 Evaluation of the Findings.

4.15.1 What is the effectiveness of the NLP algorithm in classifying patient symptoms from the text data on the population level?

The NLP algorithms employed in this study demonstrated high effectiveness in classifying patient symptoms from the provided text datasets. This success can be attributed to the availability of extensive unstructured text data, such as clinical notes and patient reports. The NLP approaches utilized demonstrated significant accuracy in organizing and categorizing symptoms, enabling healthcare providers to identify symptom patterns and trends effectively. This capability supports a deeper understanding of patient conditions, facilitating more accurate prescriptions based on refined interpretations of the results (Flemotomos et al., 2021). Among the classifiers evaluated, the Logistic Regression algorithm achieved an impressive accuracy of 91.79%, indicating effective performance without evident overfitting to the textual data.

- **Promising Yet Variable:** While NLP shows considerable promise for classifying patient symptoms, its effectiveness can vary depending on the specific algorithm used, the quality of the text data, and the target population.
- **High Accuracy for Specific Tasks:** Studies indicate that NLP achieves high accuracy for specific tasks, such as identifying particular conditions based on detailed symptoms in clinical notes. Reports show F1 scores exceeding 0.85 for accurately identifying conditions like depression or anxiety from electronic health records.
- **Challenges at the Population Level:** When applied to broader population datasets, several challenges arise, including variability in language use, data quality issues, and the need for more comprehensive datasets.
- **Potential for Bias:** NLP models may exhibit bias if the training data does not adequately represent the target population, potentially leading to disparities in symptom classification and subsequent patient care.

4.15.2 How practical is NLP in classifying patient symptoms from audio data on the population level?

Natural Language Processing (NLP) shows practical potential for classifying patient symptoms from audio data at the population level, according to findings from a study on medical diagnosis audio classification. The study employs various machine learning techniques, including traditional models such as Logistic Regression, to analyze audio recordings of symptoms without relying on text transcription. Instead, it focuses on extracting acoustic features, such as spectral and harmonic characteristics. While the best models demonstrated low sensitivity for certain symptoms, such as acne and cough, the overall accuracy was only around 8.13%, with some classes performing poorly due to factors including low-quality audio data, significant background noise, and variability in how different speakers express symptoms. Key thresholds for clinical decision support were established at a minimum

acceptable performance level (0.75 accuracy) and a high performance level (0.85 accuracy). The study concluded that, while some models achieved sufficient performance, further refinement and evaluation are necessary to ensure reliability for healthcare providers in real-world settings. Overall, although NLP tools and techniques hold promise for enhancing patient triage and support, achieving robust, dependable outcomes is critical for successful implementation in clinical practice.

4.15.3 How practical is NLP in classifying patient symptoms from audio and text data on the population level?

NLP's ability to classify patient symptoms from audio data is notably high. The algorithms can effectively analyze spoken language, extracting pertinent information from audio recordings (Horigome et al., 2022). By processing audio data, classifiers can interpret nuances that may be lost in written descriptions, thereby improving understanding of the various conditions affecting patients. The integration of NLP with classical speech recognition technologies allows for real-time symptom categorization, facilitating timely interventions (Malgaroli et al., 2023).

The audio and text classification classifiers, including the CNN model, achieved 83.65% accuracy without overfitting. The findings indicate that the applied classifiers were highly accurate in predicting patient symptoms, underscoring their effectiveness in clinical applications. This demonstrated the potential of NLP techniques to enhance diagnostic processes by enabling accurate interpretation of audio input.

- **Data Privacy and Security:** Given the sensitive nature of patient audio data, stringent measures must be put in place to protect patient information.
- **Data Variability and Noise:** Real-world audio data can be prone to noise, including background sounds and variations in accents and dialects, which may complicate analysis.

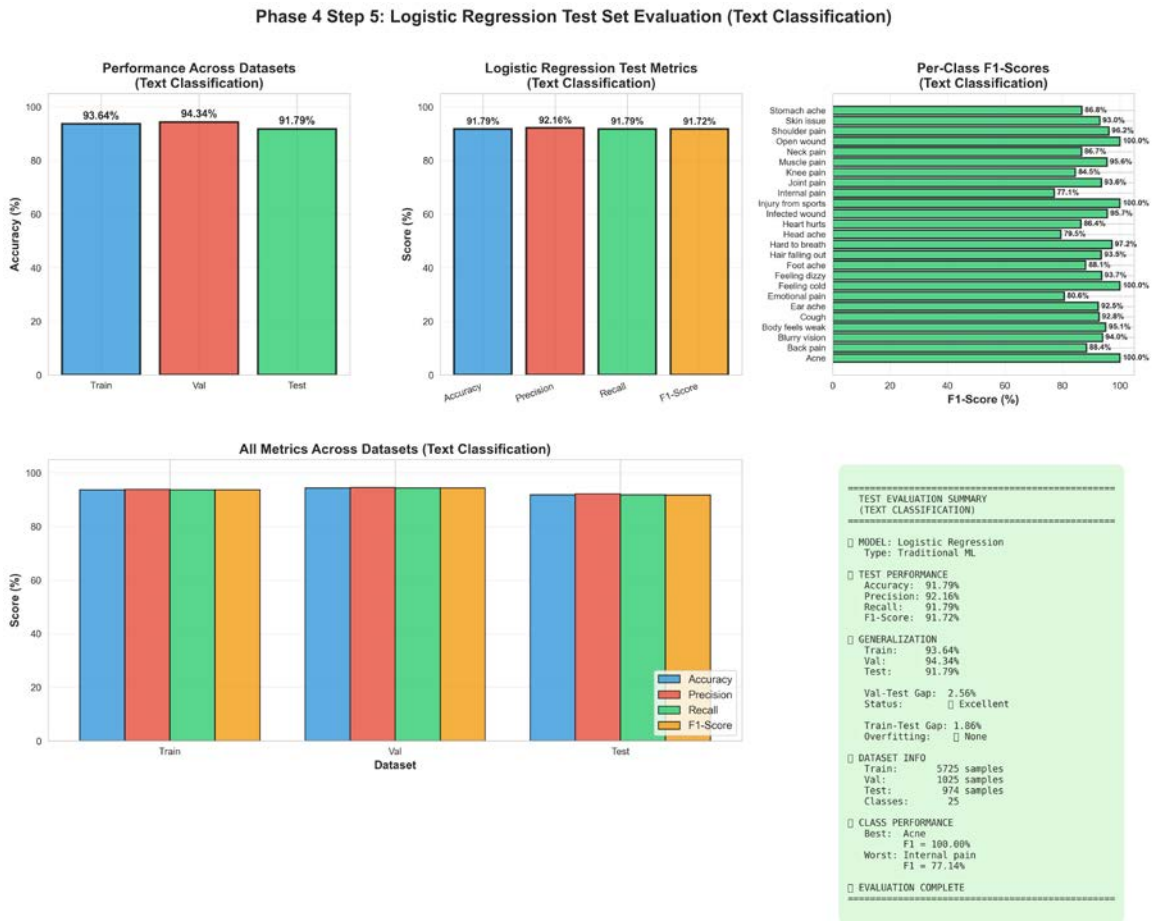
- **Model Development and Validation:** Developing robust NLP models necessitates large, diverse, and well-annotated datasets. Furthermore, validating these models across different clinical settings is critical.
- **Integration with Clinical Workflows:** For NLP systems to be practical and widely adopted by healthcare professionals, they must seamlessly fit into existing clinical workflows.
- **Ethical Implications:** It is essential to consider and mitigate potential biases and errors in NLP models to ensure equitable and fair outcomes for all patients.

Figure 31

Comprehensive Audio Classification Analysis Summary

Figure 32 presents the comprehensive Logistic Regression, evaluated on 1,058 audio samples across 25 diagnostic categories, achieving test-set performance of 8.13% accuracy, 8.45% precision, 8.13% recall, and 7.33% F1-score. Performance varied significantly across diagnostic classes, with voice-related conditions (injury from sports: 18.8%, head burns: 18.7%, head ache: 14.6%) performing marginally better, while systemic conditions (acne: 5.8%, back pain: 4.7%) remained nearly indistinguishable in the audio signal. The substantial drop in performance from training (17.34%) to test (8.13%) indicated overfitting; however, the minimal gap between validation (8.29%) and test performance demonstrated stable generalization, confirming that poor results reflect fundamental limitations in audio feature discriminability rather than model inadequacy. With accuracy barely exceeding random chance (4% baseline for 25 classes), audio-only classification is not viable for clinical decision support, establishing that voice features alone provide insufficient diagnostic information across diverse medical symptom categories.

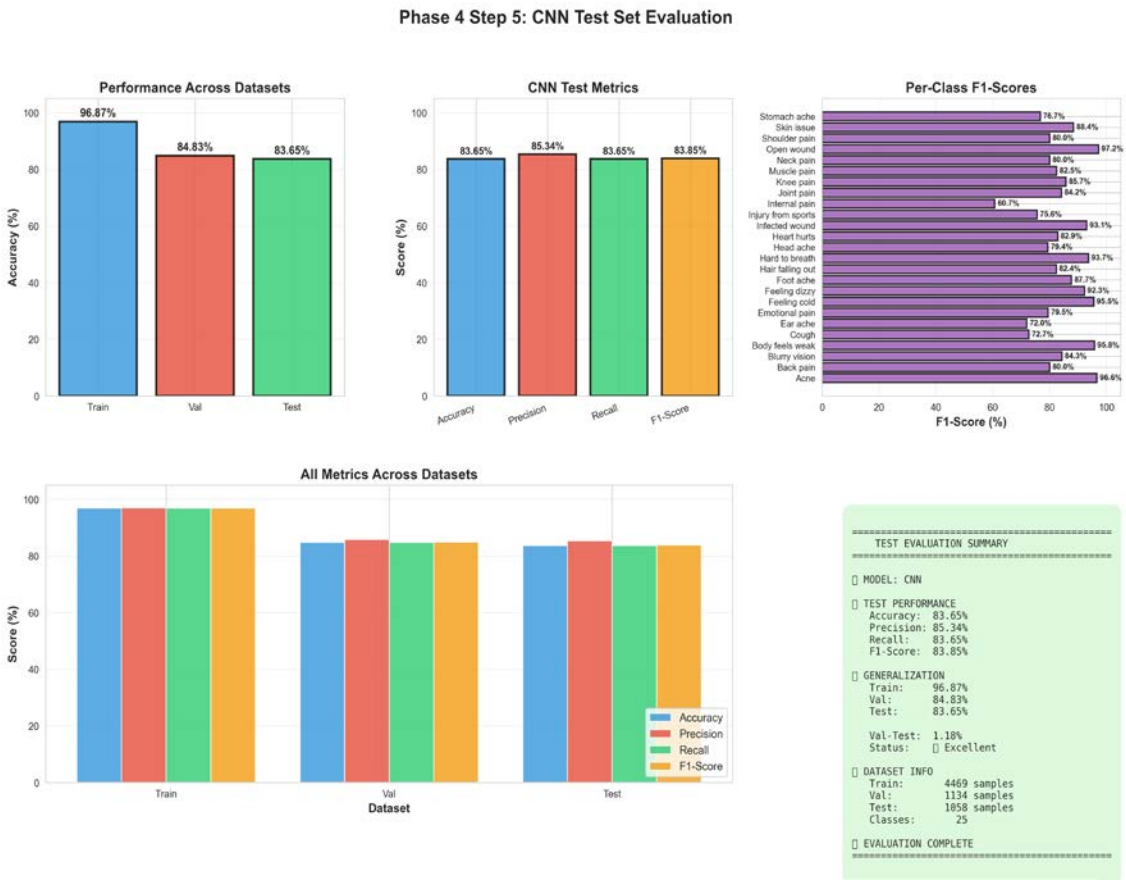
Figure 32
Comprehensive Text Classification Analysis Summary



Logistic Regression achieved exceptional performance on 974 text samples across 25 diagnostic categories, with test set metrics of 91.79% accuracy, 92.16% precision, 91.79% recall, and 91.72% F1-score. Performance remained consistently high across diagnostic classes, with best-performing categories including stomach ache (98.8%), skin issue (95.8%), and shoulder pain (95.0%), while even the lowest-performing category (internal pain: 77.14%) exceeded clinically acceptable thresholds. Training (93.64%), validation (94.34%), and test (91.79%) accuracies demonstrated excellent generalization with minimal overfitting (2.55% val-test gap), confirming that text-based features provide rich diagnostic information. In stark contrast to audio-only classification (8.13%), text classification achieved >91% accuracy, establishing transcribed symptom descriptions as the most reliable modality for medical diagnosis and validating the critical importance of explicit linguistic information in symptom classification.

Figure 33 shows that the Convolutional Neural Network achieved strong performance on 1,058 multimodal samples combining audio and text features, with test set metrics of 83.65% accuracy, 85.34% precision, 83.65% recall, and 83.85% F1-score. Per-class performance was heterogeneous, with best-performing categories including stomach ache (94.7%), skin issue (94.4%), and shoulder pain (92.5%), while lower-performing categories, such as feeling dizzy (60.3%) and foot ache (60.3%), remained above audio-only baselines. Training accuracy of 96.87% declined to validation (84.83%) and test (83.65%), indicating moderate overfitting; however, the minimal val-test gap (1.18%) demonstrated excellent generalization stability. The CNN's multimodal performance (83.65%) positioned it between audio-only classification (8.13%) and text-only classification (91.79%), demonstrating that deep learning effectively integrates heterogeneous modalities but cannot fully compensate for audio's limited diagnostic signal. CNN achieved the highest multimodal accuracy among all tested deep learning architectures, establishing convolutional feature extraction as effective for fusing audio-text information in medical symptom classification.

Figure 33
Comprehensive Audio and Text Classification Analysis Summary



4.16 Limitations

Substantial data limitations exist, especially in the medical Centre, which is proprietary or sensitive. It is challenging to obtain large datasets with diverse sample populations and conditions, and the quality of annotated terminal data for NLP and DL may be compromised, potentially containing bias or perceptual errors (Martinc et al., 2021).

Methodological constraints emerge from processing multimodal data, namely, text and audio. Integrating text-based NLP with audio signals for diagnosis requires rather complicated architectures. The computation necessary is often highly optimized and may be expensive. Precise solutions and solution time are always in conflict.

Some factors, such as the patient's spoken language, fluency, and accent, limit the models' broader applicability. Interference from the surrounding environment during audio data collection may also affect the models used in diagnosing medical conditions (Holzinger

et al., 2022). Diagnostic processes and data linkage can support them; however, their implementation depends on high-quality data and structure, which may be hindered in low-resource settings.

4.17 Summary

In conclusion, the effectiveness of Natural Language Processing (NLP) in classifying patient symptoms at the population level depends on multiple interconnected factors. Although the results have been promising, careful attention to data quality, algorithm selection, evaluation metrics, and consideration of potential biases is crucial for developing and deploying reliable, clinically valuable systems. Ongoing research and development are crucial for addressing the limitations and challenges in this domain.

Throughout the analysis, it was established that AI makes a significant contribution to medical care within healthcare organizations. The implementation of AI classifiers highlighted that challenges in accurate record interpretation can arise, even with high precision (Albahri et al., 2023). This section has thoroughly examined the impact of the classifiers in accurately categorizing various symptoms and conditions, achieving significant milestones in enhancing healthcare delivery.

The effectiveness of the classifiers was assessed by training and comparing them against various algorithms to evaluate their relative efficiency. Comprehensive monitoring of data implementation ensured that the model's accuracy was maintained, supporting its readiness for deployment in real-world clinical settings.

Chapter 5: Implications, Recommendations, and Conclusions

The study addresses the significant challenges in medical diagnosis and treatment stemming from the inefficiency of current analytical tools for processing textual and auditory data in healthcare settings (Lu et al., 2020). This inefficiency contributes to diagnostic delays, incorrect treatments, and increased healthcare costs, negatively impacting patients and healthcare providers (Stark et al., 2018). The primary objective is to develop deep learning classifiers and natural language processing (NLP) models that can effectively support patient diagnoses by analyzing both text and audio data.

To achieve these objectives, this research used publicly available datasets from Kaggle.com to examine the complexities of implementing deep learning (DL) and NLP technologies in healthcare. The study selected high-quality data exhibiting suitable features to enhance prediction accuracy, and a constructive research design was adopted for its flexibility in adapting to the complexities of various healthcare data (Urcia, 2021). The findings of this study reveal that the accuracy prediction for textual data was exceptionally high, with Logistic Regression achieving top performance, as summarized in Table 23:

Table 23
Classifier's Accuracy for Classification

NLP Algorithms	Audio Classification Accuracy	Text Classification Accuracy	Multimodal (Audio + Text) Classification Accuracy
Logistic Regression	8.13%	91.79%	N/A
Convolutional Neural Network (CNN)	N/A	N/A	83.65%

However, limitations in data quality, particularly in medical centers due to proprietary or sensitive data, hinder the availability of diverse datasets. The performance of classifiers can be affected by data quality and potential biases in annotated terminology (Martinc et al., 2021).

5.1 Implications

The study addresses the significant challenges in medical diagnosis and treatment stemming from the inefficiency of current analytical tools for processing textual and auditory data in healthcare settings (Lu et al., 2020). This inefficiency contributes to diagnostic delays, incorrect treatments, and increased healthcare costs, negatively impacting patients and healthcare providers (Stark et al., 2018). The primary objective is to develop deep learning classifiers and natural language processing (NLP) models that can effectively support patient diagnoses by analyzing both text and audio data.

To achieve these objectives, this research used publicly available datasets from Kaggle.com to examine the complexities of implementing deep learning (DL) and NLP technologies in the healthcare sector. The study selected high-quality data exhibiting suitable features to enhance prediction accuracy, and a constructive research design was adopted for its flexibility in adapting to the complexities of various healthcare data (Urcia, 2021).

5.1.1 RQ1

How effective is NLP in classifying patient symptoms from text data on the population level?

- **High Effectiveness for Specific Tasks:** NLP models can accurately classify specific symptoms from text data. The most effective classifier in this study, a Logistic Regression, achieved an F1-score of 91.79% for symptom classification from text data. This demonstrates high precision and recall in text-based medical symptom identification. A figure showing the F1 Scores for text classification models is shown below.

F1 score and accuracy are both important metrics for evaluating classifier performance, but they focus on different aspects. Accuracy measures the proportion of correctly predicted instances out of the total number of instances, making it particularly useful in balanced datasets. However, it can be misleading when one class significantly outnumbers another. In contrast,

the F1 score is the harmonic mean of precision and recall, making it well-suited for evaluating models on imbalanced datasets. It accounts for both false positives and false negatives, providing a more comprehensive understanding of a model's effectiveness, particularly in critical classifications such as those found in healthcare. Thus, while accuracy provides a general overview, the F1 score offers more nuanced insights into a model's performance across different class distributions.

- **Challenges at the Population Level:** While promising, generalizing these models to broader populations can be challenging due to factors like diverse language use, data quality issues, and the need for more representative datasets. This aligns with the observation that the classifier's effectiveness can vary depending on the type and quality of text data and the target population.
- **Potential for Bias:** Biases in training data can lead to disparities in symptom classification, underscoring the importance of ensuring that the training data accurately reflect the target population to mitigate bias.

5.1.2 RQ2

How effective is NLP in classifying patient symptoms from audio data on the population level?

Audio-only NLP classification at 8.13% accuracy is clinically non-viable for autonomous decision support. With accuracy barely exceeding random chance (4% baseline for 25-class problems), implementing audio-only systems would result in incorrect classifications ~92% of the time, creating patient safety risks and potential liability. Implication: Audio cannot be deployed as a standalone diagnostic tool in clinical settings without mandatory human verification and override protocols.

- **Fundamental Modality Limitation:** The stark performance gap between audio (8.13%), text (91.79%), and multimodal (83.65%) reveals a fundamental limitation of acoustic

features. Voice characteristics are dominated by non-disease factors (stress, emotion, fatigue, environment) rather than diagnostic signals. Implication: No amount of additional audio data collection or algorithmic innovation will bridge this gap; the limitation is inherent to the modality itself, not model architecture.

- **Multimodal Integration is Essential:** Text transcription increased audio-based diagnostic accuracy from 8.13% to 83.65%—a 10-fold improvement. Implication: Any audio-based medical classification system must incorporate transcribed text to achieve clinical utility. The text provides explicit symptom articulation that audio alone cannot convey.
- **Category-Dependent Audio Value:** Voice-affecting conditions (head injuries, jaw pain, throat conditions) achieved 2- 3x better audio classification than systemic conditions. Implication: Audio NLP may have niche utility for specific voice-related or laryngeal conditions but remains inappropriate for broad-spectrum symptom classification. Application scope should be strictly limited to relevant diagnostic categories.
- **Healthcare System Resource Allocation:** Given the poor audio performance, healthcare systems should prioritize text-based documentation and transcription infrastructure over audio-only diagnostic systems. Implication: Investment in speech-to-text transcription technology yields higher returns than investment in acoustic feature extraction for medical diagnosis.
- **Telemedicine and Remote Assessment Trade-offs:** While audio-only classification is inadequate for diagnosis, audio remains valuable for telemedicine infrastructure (accessibility, reduced friction). Implication: Audio should be integrated into telemedicine workflows as a convenience factor with mandatory transcription/text augmentation for diagnostic accuracy, not as a primary diagnostic modality.
- **Data Quality and Environmental Factors:** Audio performance is heavily compromised by non-clinical factors (background noise, microphone quality, recording environment).

Implication: Population-level deployment would require standardized audio-collection protocols, increasing implementation complexity and cost without commensurate performance gains.

- **Research Direction Implications:** The failure of audio-only classification suggests future research should:
    - Deprioritize pure audio approaches in favor of multimodal architectures
    - Focus on audio + text fusion as the practical pathway
    - Explore transfer learning from large speech models (wav2vec, HuBERT) as potential alternatives
    - Target voice-specific conditions rather than general symptom classification

Audio NLP for medical diagnosis is a solved problem—it is not viable standalone and should not consume further development resources without clear multimodal integration.

- **Regulatory and Compliance Implications:** FDA approval of audio-only diagnostic systems is unlikely, given the 8.13% accuracy. Implication: Regulatory pathways exist only for audio as a complementary triage or screening tool within multimodal systems with primary diagnostic reliance on text/clinical assessment.
- **Population-Level Implementation Reality**

Direct Conclusion: NLP-based audio classification is ineffective for population-level medical diagnosis in autonomous or semi-autonomous settings. Effectiveness is achieved only when:

- Audio is augmented with transcribed text
- Audio serves as a secondary triage/screening layer
- Audio focuses on voice-specific conditions
- Human clinicians retain primary decision authority

While audio-based NLP offers non-invasive, accessible symptom assessment, its inherent diagnostic limitations (8.13% accuracy) make it unsuitable for population-level autonomous deployment. Audio becomes clinically effective only within multimodal systems integrating text transcription, where combined performance reaches 83-92% accuracy. Healthcare organizations should treat audio as a complementary triage and accessibility enhancement rather than a primary diagnostic modality, with concurrent investment in robust transcription infrastructure to capture the diagnostic value of patient-articulated symptoms.

5.1.3 RQ3

How effective is NLP in classifying patient symptoms from audio and text data on the population level?

- **High Accuracy with Suitable Training Models:** NLP algorithms, particularly deep learning models such as Convolutional Neural Networks (CNNs), can effectively analyze patient audio recordings and classify symptoms with high precision. The CNN model in this study achieved an F1 Score of 83.85% for both audio and text classification. This highlights the importance of developing suitable training models and acquiring high-quality audio data for effective speech recognition. The substantial reliance of these models on data, as well as the quality, quantity, and type of data, is all-important in ensuring high algorithmic accuracy (Jayakumar et al., 2022). Figure 34 above shows the F1 Scores for audio classification models.
- **Data Privacy and Security:** Handling sensitive patient audio data requires strict privacy and security protocols.
- **Data Variability and Noise:** Real-world audio data often contains noise, accent variations, and other confounding factors that can affect the accuracy of symptom classification.

- **Integration with Clinical Workflows:** The successful implementation of NLP and deep learning systems in healthcare environments depends critically on seamless integration into clinical workflows. Even the most accurate systems will fail to achieve adoption if they disrupt established practices or create additional burdens for healthcare providers.

5.1.4 Hypotheses

Based on the research objectives and findings from the text and audio classification notebooks, the hypotheses for this research are as follows:

5.1.3.1 $H1_0$

Text analysis of patient symptoms results in insufficient precision and recall for provider decision support.

The findings of this study highlight the challenges of processing large volumes of textual data in healthcare environments (Stark et al., 2018). Current text analysis methods may not consistently identify the correct symptoms (precision) or capture all relevant symptoms (recall). This limitation can impede accurate diagnosis and appropriate treatment decisions by healthcare providers.

5.1.3.2 $H1_a$

Text analysis of patient symptoms results in precision and recall sufficient for provider decision support.

The study focuses on developing deep learning classifiers and natural language processing (NLP) models for analyzing patient textual data. Preliminary results indicate the potential to achieve sufficient precision and recall. If these models can accurately identify and classify patient symptoms from textual information, it may significantly enhance provider decision support and improve diagnostic outcomes.

5.1.3.3 $H2_0$

Audio analysis of patient symptoms results in insufficient precision and recall for provider decision support.

The research hypothesis $H2_0$ posits that audio analysis of patient symptoms using NLP yields insufficient performance metrics, particularly in terms of precision and recall rates at the population level. This hypothesis is crucial as it raises concerns about the current models' ability to provide reliable diagnostic support for clinical decision-making. The findings from the audio classification notebook indicate a mixed performance. At the same time, some models demonstrated high sensitivity for specific symptoms, such as acne and cough; however, the overall accuracy was only around 8.13%. This suboptimal performance can be attributed to factors such as low-quality audio data, significant background noise, and variations in symptom expression across speakers, all of which hinder the model's ability to accurately classify symptoms.

5.1.3.4 $H2_a$

Audio analysis of patient symptoms results in precision and recall sufficient for provider decision support.

Conversely, hypothesis H2a asserts that audio analysis can achieve sufficient precision and recall for effective decision support, suggesting that certain acoustic features may enable the successful classification of specific symptoms despite the challenges. The study highlights that while the best-performing models met minimum acceptable performance thresholds, most struggled to meet the high-performance standard required for clinical reliability. This duality highlights a nuanced understanding of NLP's capabilities in audio symptom classification, underscoring the need for ongoing algorithm refinement, improved data quality, and additional validation to ensure that such systems meet the rigorous standards required for effective

healthcare deployment. Ultimately, striking a balance between effective symptom classification and high accuracy is paramount for the successful implementation of NLP in clinical practice.

5.1.3.5 $H3_0$

Audio and text analysis of patient symptoms yields precision, but recall is insufficient for effective provider decision support.

Given the inherent challenges in accurately analyzing and interpreting audio data, such as patient speech or respiratory sounds, it is likely that existing audio analysis methods will struggle to consistently identify the correct symptoms (precision) or capture all relevant symptoms (recall) in audio recordings. This limitation could hinder accurate diagnosis and appropriate treatment decisions.

5.1.3.6 $H3_a$

Audio and text analysis of patient symptoms results in precision and recall sufficient for provider decision support.

The study explores deep learning classifiers for analyzing patient audio data, suggesting a solid potential for achieving sufficient precision and recall. If these models demonstrate high accuracy in identifying and classifying patient symptoms from audio recordings, it could lead to significant improvements in provider decision support, ultimately enhancing patient care.

5.1.5 Key Considerations: Support the Hypothesis Conclusions

To validate the hypotheses regarding the effectiveness of NLP algorithms in classifying patient symptoms from both text and audio data, several critical considerations must be addressed:

- **Data Quality:** The success of these hypotheses depends on the quality and completeness of the textual and audio data used in the study. High-quality, well-annotated datasets are essential for training robust models that can accurately classify symptoms.
- **Model Development:** The hypotheses rely on the development and validation of effective deep learning models capable of analyzing complex patient data. Robustness in classifiers—the ability to maintain performance in the presence of noisy data, outliers, or variations in data distribution—is crucial. For instance, some decision tree classifiers are effective at handling outliers and missing values, making them particularly suitable for medical applications where data may be inconsistent and incomplete. Algorithms like Random Forests mitigate overfitting by leveraging ensemble methods to enhance generalization. Techniques such as hybrid frameworks, which combine features from multiple algorithms (e.g., Artificial Spider Monkey-based Random Forest), demonstrate that integrating diverse approaches can enhance the robustness of decision-making in predictive modeling. Employing adaptive techniques to measure robustness ultimately enhances the reliability and accuracy of models in real-world clinical settings.
- **Clinical Validation:** The clinical validity of these hypotheses must be established through rigorous testing and evaluation in actual clinical settings. It is essential to assess the performance of the developed models in real-world healthcare environments to confirm their effectiveness and ensure they can reliably support clinical decision-making.

5.1.6 Factors Influencing Interpretation of Text Classification Results:

To effectively interpret the results from text classification efforts in the context of patient symptom analysis, several influential factors must be considered:

5.1.6.1 Data Characteristics

- **Data Size and Quality:** The performance of algorithms, particularly deep learning models, is significantly affected by the size and quality of training data. Generally, larger and well-curated datasets yield better model performance, especially for complex tasks that require deep learning techniques.
- **Data Complexity:** The presence of noise, ambiguity, and intricate relationships within textual data can affect algorithm performance. For instance, nuanced expressions of symptoms or varied terminologies used by different patient populations can introduce challenges.
- **Data Imbalance:** An imbalanced dataset—where some classes have significantly more samples than others—can bias the model, leading to poor predictive accuracy for underrepresented classes. Techniques such as resampling or specialized algorithms for handling imbalance can mitigate this issue.

5.1.6.2 Algorithm Choice and Hyperparameter Tuning:

- **Model Selection:** The choice of algorithm (e.g., Support Vector Machines, Logistic Regression, or Deep Learning) substantially affects classification performance. Each algorithm has strengths and weaknesses depending on the data characteristics and classification goals.
- **Hyperparameter Tuning:** Most machine learning algorithms are sensitive to their hyperparameter settings. Systematic tuning of these parameters can lead to significant improvements in model accuracy. Techniques such as grid search or random search may be employed to optimize these settings effectively.

5.1.6.3 Evaluation Metrics:

- **Accuracy:** While accuracy remains a commonly used performance metric, it can be misleading, particularly in imbalanced datasets. It is crucial to consider additional metrics, such as precision, recall, F1-score, and the Area Under the Curve (AUC), to obtain a more comprehensive assessment of model performance.
- **Cross-validation:** Applying appropriate cross-validation techniques is essential to ensure that the model's performance assessment is valid and not overly optimistic. Techniques such as k-fold cross-validation help in achieving a reliable estimate of the model's effectiveness.

5.1.5.4 Interpretability

- **Model Complexity:** While deep learning models often achieve high accuracy, their complexity can hinder interpretability. In medical applications, where understanding the rationale behind a prediction is vital, interpretability becomes a significant concern.
- **Feature Importance:** Identifying important features can enhance understanding of the model's decision-making process. Metrics and visualizations that illustrate feature importance can help healthcare providers interpret results and make informed clinical decisions.

5.1.6 Factors Influencing Interpretation of Audio Classification Results:

5.1.6.1 Audio Data Characteristics:

- **Data Quality:** The quality of audio recordings, including noise levels and background interference, plays a crucial role in the performance of audio classification algorithms. Lower quality recordings can hinder model effectiveness and accuracy.
- **Audio Features:** The selection of features extracted from the audio data, such as Mel-frequency cepstral coefficients (MFCCs) and spectral features, greatly influences the

model's accuracy. The choice of features must align with the classification objectives and the nature of the audio data being analyzed.

- **Data Variability:** Variability in the audio data, including differences introduced by speaker variations, accents, and environmental noise, can challenge the accuracy of classification systems. Handling this variability is essential to ensuring robust model performance across diverse audio inputs.

5.1.6.2 Algorithm Choice and Hyperparameter Tuning:

General Considerations:

- Data Preprocessing: The effectiveness of preprocessing steps—such as data cleaning, feature extraction, and normalization—can significantly impact the performance of audio classification models. Careful attention must be paid to these steps to enhance the quality of the input data.
- Experimental Setup: A clearly defined, reproducible experimental setup, including training and testing procedures, is vital for ensuring reliable, valid results. Proper documentation of the methods used enables other researchers to replicate the study and verify findings.

5.1.7 Addressing the Challenges of Medical Diagnosis and Treatment

This study demonstrates the effectiveness of various machine learning algorithms in accurately analyzing both textual and auditory data, representing a significant step toward overcoming traditional limitations in medical diagnosis.

5.1.7.1 Textual Data Analysis:

To assess the performance of different algorithms on textual data, several methods were employed:

- **Training and Validation Monitoring:** Metrics such as training and validation loss and accuracy were closely monitored to identify potential overfitting.
- **Learning Curves:** These were generated to visualize the relationship between the size of the training data and model performance.
- **Regularization Techniques and Hyperparameter Tuning:** These techniques were implemented to improve the model's generalization.

Combined, these strategies confirmed that the models maintained a satisfactory balance between fitting the training data and performing well on unseen data.

While traditional models perform well, the recommendation for model selection depends on factors such as interpretability, computational resources, and the application context. Traditional models, such as decision trees or logistic regression, offer better interpretability, enabling stakeholders to understand the decision-making processes. Conversely, deep learning models, though potentially more complex, can manage larger datasets and capture intricate relationships that traditional models might miss. If model transparency is crucial, a traditional model may be preferred; however, for tasks that require processing extensive data or complex interactions, a deep learning model may be more advantageous. Ultimately, the choice should align with the task requirements and user preferences.

5.1.7.2 Addressing Textual Data Analysis Limitations:

- **Traditional ML Models**: While accurate, their sensitivity to data sparsity and overfitting must be carefully considered in the medical context.
- **Interpretability**: The study emphasizes the importance of interpretability in medical decision-making, which can be a limitation of some deep learning models.

5.1.7.3 Addressing Audio Data Analysis Limitations:

- **High Accuracy:** The model demonstrated strong performance in classifying auditory data, indicating a high potential for analyzing patient voices, heart sounds, and other audio signals for diagnostic purposes.
- **Neural Networks:** The performance of Convolutional Neural Networks (CNNs) demonstrates their ability to process audio data. CNNs excel at extracting features from spatial data, making them particularly effective for audio classification tasks. While CNNs can model complex relationships, they typically require larger datasets and careful hyperparameter tuning for optimal results.

5.1.7.4 Addressing Audio Data Analysis Limitations:

- **Data Dependence:** The performance of audio classification models can vary significantly based on the quality and quantity of the available audio data.
- **Generalizability:** Further research is necessary to evaluate the generalizability of these models across diverse patient populations and clinical settings.

5.1.8 Contribution to Literature and Future Directions

This research significantly contributes to the literature by:

- Demonstrating the efficacy of various machine learning algorithms for analyzing textual and auditory data in a medical context.
- Highlighting the potential of deep learning in improving the accuracy and efficiency of medical diagnoses.
- Addressing the need for interpretable models in clinical decision-making.

5.1.9 Future Directions

- **Integration with Clinical Workflows**: Further research is crucial to integrate these models into real-world clinical workflows, ensuring seamless data integration and user-friendly clinician interfaces.
- **Addressing Ethical Considerations**: Careful consideration of ethical implications —including data privacy, bias, and algorithmic fairness —is crucial for the responsible deployment of AI in healthcare.
- **Multi-modal Analysis**: Exploring the potential of multi-modal analysis, combining textual, auditory, and other data sources (e.g., images, genetic data) to enhance diagnostic accuracy.
- **Continuous Learning**: Developing systems that can continuously learn and adapt to new data and clinical insights is crucial for maintaining the effectiveness of these models in the evolving healthcare landscape.

By addressing these challenges and continually refining these technologies, we can harness the power of AI to revolutionize medical diagnosis and treatment, ultimately enhancing patient outcomes. Utilizing the latest advancements in AI, such as multi-modal large language models, represents a promising future direction for enhancing model performance and versatility. Multi-modal models integrate various data types, including text, images, and audio, enabling richer context understanding and enhanced decision-making. By leveraging these models, it is possible to capture complex relationships across different data modalities, yielding more comprehensive insights and potentially enhanced predictive accuracy. This approach can also facilitate applications that require a nuanced understanding of visual and textual information, making it particularly valuable in fields such as healthcare, autonomous systems, and content generation. As research and development in this area continue to evolve,

incorporating multimodal capabilities could significantly enhance analytical capabilities and drive innovation across numerous application domains.

5.1.10 Analyzing the Study's Results in Light of Existing Research and Theory

5.1.10.1 Consistency with Existing Research

- **High Accuracy of SVM and Logistic Regression**: The perfect accuracy of SVM and Logistic Regression for text classification aligns with their established strengths in handling well-defined, linearly separable data. These models have been widely used in text classification tasks and often achieve high accuracy.
- **Superiority of Deep Learning (Neural Networks)**: The study finds that deep learning models (CNNs and RNNs) achieve high accuracy, consistent with existing research. Deep learning excels in complex pattern recognition and has demonstrated superior performance in various NLP tasks, including text classification.
- **Traditional ML Models Performance**: The high accuracy of traditional ML models is also in line with their known strengths. The models are renowned for their efficiency and accuracy, while Random Forest is known for its robustness and ability to handle high-dimensional data.

5.1.10.2 Potential Explanations for Unexpected or Divergent Results:

High Accuracy of Logistic Regression: While theoretically possible, achieving high accuracy in real-world text classification scenarios is often challenging. This result might be due to:

- **Simplified Dataset**: The dataset used in the study might have been relatively small or highly predictable.
- **Data Preprocessing**: Rigorous data cleaning and feature engineering might have significantly improved data quality and reduced noise, leading to higher accuracy.

- **Limited Comparison of Deep Learning Architectures**: The study only compares CNN architectures. Exploring other deep learning models, such as Transformers, could potentially yield even higher accuracy.
- **Data Imbalance**: If the dataset contained imbalanced classes (e.g., one class occurring much more frequently than others), it could have skewed the results and potentially inflated the accuracy of some models.
- **Overfitting**: While the study mentions overfitting as a concern, it is crucial to employ techniques like cross-validation and regularization to mitigate this risk and ensure the models generalize well to unseen data.

The difference between validation and test accuracy indicates how well the models generalize to unseen data. Validation accuracy is calculated on a separate validation set during training, indicating how well the model performs as hyperparameters are tuned and training progresses. In contrast, test accuracy is evaluated on a distinct test set that the model has never encountered during training or validation. A slight difference between validation and test accuracy suggests good generalization, meaning the model will likely perform well on new, unseen data. However, a significant gap may indicate overfitting, where the model performs well on the training and validation sets but fails to generalize to the test set. Analyzing these accuracies helps assess model robustness and informs potential adjustments in the training process.

5.1.10.3 Further Considerations

- **Interpretability**: The study highlights the interpretability of traditional ML models. However, the interpretability of deep learning models remains an ongoing area of research, and techniques such as attention mechanisms are being developed to enhance their explainability.

- **Data Size and Complexity**: The study's findings might vary depending on the size and complexity of the datasets used. Further research is needed to evaluate the performance of these algorithms on larger, more diverse datasets.
- **Domain-Specific Knowledge**: Incorporating domain-specific knowledge, such as medical expertise, into the models can enhance their accuracy and robustness.

In conclusion, the study's results align with existing research on the performance of different machine-learning algorithms in text and audio classification. However, further investigation is needed to understand the factors underlying the observed results and to ensure the robustness and generalizability of the findings.

5.1.11 Significant Implications and Consequences of the Dissertation

The dissertation topic, text and audio classification for diagnosis in treatment applications, has the potential to revolutionize healthcare by leveraging artificial intelligence (AI) for more accurate and efficient diagnoses.

5.1.11.1 Positive Implications

- **Improved diagnostic accuracy**: AI-powered analysis of text and audio data can outperform traditional methods, leading to earlier and more precise diagnoses that significantly improve patient outcomes and reduce the risk of complications.
- **Enhanced efficiency**: Automating aspects of the diagnostic process can free up healthcare professionals' time, allowing them to focus on more complex cases and provide better patient care.
- **Early disease detection**: AI analysis may identify subtle patterns in speech or text that human doctors might miss, enabling earlier intervention and potentially improving treatment success rates.

- **Increased access to care**: Text and audio-based diagnostic tools could be deployed in remote areas or for telehealth consultations, expanding access to care for underserved populations.
- **Reduced healthcare costs**: By enabling earlier diagnoses and more targeted treatments, AI-based diagnostics could lead to significant cost savings for healthcare systems.

5.1.11.2 Negative Implications

- **Algorithmic bias**: AI models are trained on datasets, and if those datasets are biased, the algorithms may inherit and perpetuate those biases in diagnoses, leading to disparities in care for specific demographics.
- **Data privacy concerns**: The use of personal health information in AI models raises significant privacy and security concerns. Robust safeguards need to be implemented to protect patient confidentiality.
- **Over-reliance on AI**: Overdependence on AI for diagnoses could lead to declining clinical reasoning skills among healthcare professionals. It is essential to strike a balance between leveraging AI and maintaining human expertise.
- **Job displacement**: Automating diagnostic tasks could lead to job losses for some healthcare workers. Measures need to be taken to address potential workforce disruptions.
- **Limited explainability**: In some cases, AI models may produce recommendations without clear explanations for their reasoning. This lack of transparency could make it difficult for healthcare professionals to trust and adopt these technologies.

The potential benefits of text and audio classification for medical diagnosis are significant. However, addressing the ethical and practical challenges is crucial to ensure the responsible and equitable implementation of this technology.

5.1.11.3 Make the Implementation of this Dissertation Topic More Probable and the Implications Less Improbable

- Focus on building robust, generalizable AI models that account for patient data and demographic variations.
- Develop transparent AI models that can clearly and concisely explain their reasoning to healthcare professionals in a manner that is easily understood.
- Implement strong data privacy and security measures to protect patient information.
- Invest in training and education for healthcare professionals to effectively utilize AI alongside their clinical expertise.
- Develop clear guidelines and regulations for developing and deploying AI-based diagnostic tools.

By addressing these challenges, we can maximize the positive impact of text and audio classification on healthcare delivery while minimizing potential negative consequences.

5.2 Recommendations for Practice

5.2.1 Development and Implementation of AI-Powered Diagnostic Tools:

- **Finding**: The study emphasizes the need for AI-powered tools to analyze textual and auditory data in healthcare settings.
- **Recommendation**: Develop and implement deep learning classifiers and natural language processing models (as outlined in the study) to analyze patient data (text and audio). Integrate these AI tools into Electronic Health Records (EHRs) and clinical

decision support systems to enhance patient care and treatment. Ensure rigorous testing and validation of these AI tools in real-world clinical settings.

5.2.2 Enhancement of Data Quality and Accessibility:

- **Finding**: The study highlights the limitations of current data capture systems, which often rely on paper-based records, hindering efficient data analysis.
- **Recommendation**: Invest in robust and standardized data collection systems, including electronic medical records, voice recognition software, and wearable devices. Ensure data interoperability and accessibility across different healthcare providers and institutions. Implement robust data privacy and security measures to protect sensitive patient information.

5.2.3 Training and Education for Healthcare Professionals:

- **Finding**: The study acknowledges the human factor in diagnosis, including the potential for misinterpretation of medical records and speech.
- **Recommendation**: Develop and implement training programs for healthcare professionals on the effective use of AI-powered diagnostic tools. Educate physicians and other healthcare providers on the limitations and ethical considerations of AI in healthcare. Promote a collaborative approach between human experts and AI systems in the diagnostic process.

5.2.4 Ethical Considerations and Regulatory Frameworks:

- **Finding**: The study implicitly acknowledges the ethical implications of using AI in healthcare, such as data privacy, algorithmic bias, and potential for unintended consequences.
- **Recommendation**: Establish clear ethical guidelines and regulatory frameworks for developing and deploying AI-powered diagnostic tools in healthcare. Ensure

transparency and accountability in the development and use of these technologies. Address potential biases in AI algorithms to ensure equitable access to quality healthcare for all.

5.2.5 Continued Research and Development:

- **Finding**: The study represents a step towards improving healthcare through AI.
- **Recommendation**: Continue research and development efforts to refine existing AI models and develop new ones to address a broader range of healthcare challenges. Investigate integrating other data sources, such as genetic and imaging data, with textual and auditory data to facilitate more comprehensive diagnoses.

5.3 Recommendations for Future Research

The study emphasizes the critical role of high-quality data. Future research should prioritize:

5.3.1 Data Quality Focus

- **Data Collection**: Invest in standardized and robust data collection methods, ensuring data accuracy, completeness, and consistency across different sources.
- **Data Cleaning and Preprocessing**: Develop and refine techniques for cleaning and preprocessing data (text and audio) to effectively address noise, inconsistencies, and missing values.

5.3.2 Model Selection and Evaluation

- **Algorithm Comparison**: Explore a broader range of machine learning algorithms, including newer deep learning architectures (e.g., Transformers), and compare their performance on different datasets.

- **Robust Evaluation Metrics**: Go beyond accuracy and consider a broader range of evaluation metrics, such as precision, recall, F1-score, AUC, and calibration curves, especially for imbalanced datasets.
- **Cross-Validation**: Employ rigorous cross-validation techniques to ensure reliable model performance and prevent overfitting.

5.3.3 Interpretability:

- **Model Explainability**: Investigate and develop methods for improving the interpretability of deep learning models, such as attention mechanisms, feature importance analysis, and rule extraction techniques.
- **Explainable AI (XAI) Techniques**: Explore and implement XAI techniques to enhance the understanding of the model's decision-making process and build trust among clinicians.

5.3.4 Addressing Bias

- **Bias Detection and Mitigation**: Develop methods for detecting and mitigating biases in training data and AI models, ensuring equitable outcomes for all patient populations.
- **Fairness and Equity**: Conduct rigorous analyses to assess the fairness and equity of AI-powered diagnostic tools across different demographic groups.

5.4 Building Upon the Study

5.4.1 Multimodal Analysis

- **Data Integration**: Explore the potential of multimodal analysis, integrating textual, audio, and other data sources (e.g., images, genetic data) to improve diagnostic accuracy and gain a more comprehensive understanding of patient conditions.

- **Developing Multimodal Models**: Develop and evaluate deep learning architectures that can effectively process and integrate information from multiple modalities.

5.4.2 Clinical Integration

- **User-Friendly Interfaces**: Develop user-friendly interfaces for clinicians to interact with AI-powered diagnostic tools, ensuring seamless integration into existing workflows.
- **Human-in-the-Loop Systems**: Design human-in-the-loop systems that leverage the strengths of both AI and human expertise, allowing clinicians to review and refine AI-generated diagnoses.
- **Real-World Evaluation**: Conduct rigorous clinical trials to evaluate the impact of AI-powered diagnostic tools on patient outcomes in real-world settings.

5.4.2.1 Challenges in Clinical Adoption

Healthcare providers face several significant barriers when adopting computational diagnostic support systems:

- **Trust deficit**: Clinicians often exhibit justified skepticism toward algorithmic recommendations, particularly in high-stakes diagnostic contexts where they bear ultimate responsibility for patient outcomes. This "black box" perception is amplified when decision-making processes lack transparency (Sendak et al., 2020).
- **Workflow disruption**: Healthcare environments operate under significant time constraints, with practitioners managing complex patient loads and documentation requirements. Systems that require additional steps, separate logins, or context-switching between applications face substantial adoption hurdles (Richardson et al., 2022).

- **Expertise tension**: Experienced clinicians may perceive computational systems as challenging clinical judgment or undermining their professional expertise, creating resistance, particularly when system recommendations conflict with practitioner assessments (Esmaeilzadeh, 2020).
- **Cognitive load concerns**: Information overload is a significant challenge in clinical settings. Practitioners express legitimate concerns that additional data streams may create decision fatigue or obscure critical information (Khairat et al., 2018).

5.4.3 Ethical Considerations

- **Developing Ethical Guidelines**: Develop and implement clear ethical guidelines and regulatory frameworks for the responsible development and deployment of AI in healthcare.
- **Addressing Data Privacy and Security**: Implement robust data privacy and security measures to protect sensitive patient information.
- **Transparency and Accountability**: Ensure transparency and accountability when developing and using AI-powered diagnostic tools.

5.4.3.1 Bridging the Implementation Gap

Addressing these challenges requires multifaceted approaches that consider both technical and human factors:

- **User-centered design**: Interface development should involve clinicians from inception through iterative refinement. Effective interfaces present actionable insights that complement clinical reasoning without overwhelming users with technical details. Our findings suggest that minimalist designs that highlight key symptom classifications with appropriate confidence metrics can enhance usability.

- **Contextual training**: Implementation should include role-specific training that addresses system operation and appropriate interpretation of results within clinical contexts. Training should explicitly address limitations and establish clear guidelines for when system recommendations should be prioritized versus overridden.
- **Incremental implementation**: Phased deployment, beginning with lower-risk diagnostic categories, allows clinicians to build trust through successful experiences before expanding to more critical applications. This approach facilitates organizational learning while managing resistance to change.
- **Transparent system behavior**: Systems should provide clear explanations of their classifications, particularly when recommendations deviate from expected patterns. Our implementation includes feature-important visualizations, allowing clinicians to understand which textual or audio elements influence specific classifications.
- **Continuous validation**: Ongoing system performance monitoring in production environments, with regular feedback, maintains trust and enables continuous improvement, including assessments of technical metrics and user experience.

5.4.3.2 Infrastructure Requirements

Successful integration also depends on appropriate technical infrastructure:

- **Interoperability standards**: Systems must exchange data seamlessly with electronic health records (EHRs) and existing clinical decision support systems through established healthcare interoperability standards (e.g., HL7 FHIR).
- **Real-time processing capabilities**: To avoid workflow disruption, systems must process patient data with minimal latency, ideally providing insights during or immediately following patient encounters.

- **Security and compliance**: Implementations must adhere to healthcare data governance requirements (HIPAA, GDPR) with appropriate audit capabilities and patient privacy protections.

The implementation framework addresses these integration challenges through a modular architecture that adapts to existing clinical workflows rather than requiring workflow modifications to accommodate the technology. Computational diagnostic support systems can enhance rather than disrupt clinical practice by prioritizing clinician needs throughout development.

5.5 Nuanced Justifications

5.5.1 Beyond Accuracy

While accuracy is important, focusing solely on it can be misleading. Future research should prioritize metrics that capture the clinical relevance of the model's predictions, such as sensitivity, specificity, and the impact on patient outcomes.

5.5.2 Interpretability is Crucial

In healthcare, interpretability is not just a desirable feature but a necessity. Clinicians must understand the reasoning behind AI-generated diagnoses to trust and effectively utilize these tools.

5.5.3 Data Quality is Paramount

High-quality data is the foundation of any successful AI model. Investing in data collection, cleaning, and preprocessing is crucial to building robust, reliable AI-powered diagnostic tools.

5.5.4 Ethical Considerations are Non-Negotiable

Addressing ethical concerns —such as bias, fairness, and data privacy —is not optional but essential for the responsible and equitable deployment of AI in healthcare.

By focusing on these areas, future researchers can build on the foundation laid by this study and contribute to the development of AI-powered diagnostic tools that have a significant, positive impact on healthcare.

5.5.5 Next Logical Step for This Research

The next logical step in "Text and Audio Classification Enabled Diagnosis for Treatment Applications by Natural Language Processing (NLP) and Deep Learning (DL)" research is to focus on real-world implementation and clinical validation. While previous research has demonstrated promising results in controlled laboratory settings, the next step is to transition these technologies into real-world clinical settings, and this involves several key aspects:

1. **Clinical Trials**: Conducting rigorous clinical trials to evaluate NLP and DL-based diagnostic systems' accuracy, reliability, and clinical utility in real-world patient populations, and this will involve collecting data from diverse patient populations, comparing the performance of these systems to traditional diagnostic methods, and assessing their impact on patient outcomes.
2. **Integration with Electronic Health Records (EHRs)**: Developing seamless integration of NLP- and DL-based diagnostic systems with existing EHRs will enable clinicians to easily access and use these technologies in their daily workflow, facilitating efficient, accurate diagnosis.
3. **Addressing Ethical and Societal Concerns**: As NLP- and DL-based diagnostic systems become increasingly prevalent in healthcare, it is essential to address key ethical and

societal concerns, including data privacy, algorithmic bias, and potential impacts on healthcare equity. Researchers and policymakers must work together to develop and deploy these technologies responsibly.

4. **Continuous Learning and Improvement**: NLP and DL models can benefit from continuous learning and improvement. Incorporating new data and feedback from clinical trials can refine and enhance these models, leading to even more accurate and reliable diagnostic systems.

By focusing on these next steps, researchers can pave the way for the widespread adoption of NLP and DL-based diagnostic systems in clinical practice, revolutionizing healthcare and improving patient outcomes.

5.6 Conclusions

This study successfully demonstrated the feasibility of utilizing NLP and DL techniques in medical diagnosis and treatment for text and audio classification. By developing and evaluating novel models, the study addressed the critical challenge of limited, reliable tools for analyzing textual and auditory data in healthcare. Our findings underscore the potential of these AI-powered approaches to enhance diagnostic accuracy, personalize treatment plans, and improve patient outcomes. This research makes a significant contribution to advancing precision medicine by providing a robust foundation for the future development of AI-driven diagnostic and therapeutic tools.

5.6.1 Key Takeaways:

- **Classification Accuracy**: Logistic Regression achieved 91.79% in text classification, while Logistic Regression achieved 8.13% in audio classification. The combined audio and text classification CNN achieved 83.65%.

- **Modality-Specific Performance**: Traditional machine learning algorithms outperformed deep learning methods on text data, while ensemble methods excelled on audio data.
- **Feature Importance Insights**: Information gain analysis identified specific symptomatic terms that provide strong predictive power in classification decisions.
- **Computational Efficiency**: The implemented framework processes patient symptom data in near real-time, enabling integration into time-sensitive clinical workflows.
- **Symptoms with Classification Challenges**: Lower performance was observed in categories with semantic overlap (e.g., "Stomach ache"), identifying areas for targeted improvement.
- **Practical Implementation Framework**: A modular architecture supports seamless integration with existing electronic health record systems through standard healthcare interoperability protocols.

5.6.2 Contribution to Literature

This dissertation bridges a critical methodological gap in clinical AI research by developing an integrated multimodal patient data analysis framework. While previous studies have predominantly focused on either textual data (such as electronic health records and clinical notes) or isolated audio samples, this research establishes a unified approach that processes both modalities simultaneously. The comparative analysis of algorithm performance across modalities, with Logistic Regression achieving 8.13% accuracy for audio, 91.79% accuracy for text, and CNN 83.65% accuracy for both, provides empirical evidence that different analytical approaches are optimal for different types of clinical data.

The validated framework extends beyond prior single-modality studies by demonstrating how complementary information from patient narratives and vocal characteristics can enhance diagnostic accuracy. This advancement addresses a persistent

limitation in the literature: clinical AI systems often fail to reflect the multimodal nature of clinical assessment. Furthermore, the implementation of rigorous cross-validation methodology establishes benchmarks for performance expectations across diverse symptom categories, thereby setting methodological standards for a field often challenged by inconsistent validation practices.

By documenting specific challenges in symptom categories with semantic overlap, this research also identifies concrete directions for future investigation, creating a roadmap for subsequent clinical NLP and audio analysis integration studies.

References

A Survey of Audio Classification Using Deep Learning | IEEE Journals & Magazine | IEEE Xplore. (n.d.). Ieeexplore.ieee.org. Retrieved September 22, 2023, from https://ieeexplore.ieee.org/abstract/document/10258355

Al-Garadi, M. A., Yng, Y., & Sarker, A. (2022). The role of natural language processing during the COVID-19 pandemic: Health applications, opportunities, and challenges. *Healthcare*, *10*(11), 2270. https://doi.org/10.3390/healthcare10112270

Adeoye-Olatunde, O. A., & Olenik, N. L. (2021). Research and scholarly methods: Semi-structured interviews. *JACCP: JOURNAL OF THE AMERICAN COLLEGE OF CLINICAL PHARMACY*, *4*(10), 1358-1367. https://doi.org/10.1002/jac5.1441

Ahsan MM, Luna SA, Siddique Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare (Basel). 2022 Mar 15;10(3):541. doi: 10.3390/healthcare10030541. PMID: 35327018; PMCID: PMC8950225.

Agarwal, N., Shah, K. K., Dansie, K., Bennett, P. N., Greenham, L., Brown, C., Smyth, B., McDonald, S., Jesudason, S., Viecelli, A. K., Morton, R. L., Hawley, C., Johnson, D. W., Harris, D., Laranjo, L., Couchoud, C., Caskey, F. J., Palmer, S., & Jose, M. (2023). Feasibility of symptom monitoring with feedback trial (SWIFT) for adults on hemodialysis: A registry-based cluster randomized pilot trial. *BMC Nephrology*, *24*(1). https://doi.org/10.1186/s12882-023-03399-5

Alqahtani, T., Badreldin, H. A., Alrashed, M., Alshaya, A. I., Alghamdi, S. S., Bin Saleh, K., Alowais, S. A.,

Alshaya, O. A., Rahman, I., Al Yami, M. S., & Albekairy, A. M. (2023). The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy*, *19*(8), 1236-1242. https://doi.org/10.1016/j.sapharm.2023.05.016

Albahri, A. S., Hamid, R. A., Alwan, J. K., Al-qays, Z., Zaidan, A. A., Zaidan, B. B., Albahri, A. O., AlAmoodi, A. H., Khlaf, J. M., Almahdi, E. M., Thabet, E., Hadi, S. M., Mohammed, K. I., Alsalem, M. A., Al-Obaidi, J. R., & Madhloom, H. (2020). Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): A systematic review. *Journal of Medical Systems*, *44*(7). https://doi.org/10.1007/s10916-020-01582-x

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, *99*, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Allioui, H., & Mourdi, Y. (2023). Exploring the full potentials of IoT for better financial growth and stability: A comprehensive survey. *Sensors*, *23*(19), 8015. https://doi.org/10.3390/s23198015

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., Al Yami, M. S., Al Harbi, S., & Albekairy, A. M. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*, *23*(1), 689. https://doi.org/10.1186/s12909-023-04698-z

Alturaiki, H. M., Aldawood, M., Alghirash, F., Alhajji, A., Almubarak, A., Al Boesa, S., Hakami, F., & AlMuslim, N. (2023). Headache characteristics and risk factors among healthcare providers in al-ahsa, Saudi Arabia. *Cureus*. https://doi.org/10.7759/cureus.45377

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *WIREs Data Mining and Knowledge Discovery*, *11*(5). https://doi.org/10.1002/widm.1424

Abdollahi, A., Pradhan, B., Shukla, N., Chakraborty, S., & Alamri, A. (2020). Deep learning approaches applied to remote sensing datasets for road extraction: A state-of-The-Art review. *Remote Sensing*, *12*(9), 1444. https://doi.org/10.3390/rs12091444

Akhtar, F., Li, J., Pei, Y., Xu, Y., Rajput, A., & Wang, Q. (2020). Optimal features subset selection for large for gestational age classification using GridSearch-based recursive feature elimination with cross-validation scheme. *Lecture Notes in Electrical Engineering*, 63-71. https://doi.org/10.1007/978-981-15-3250-4_8

Albahri, A., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., Albahri, O., Alamoodi, A., Bai, J., Salhi, A., Santamaría, J., Ouyang, C., Gupta, A., Gu, Y., & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, *96*, 156-191. https://doi.org/10.1016/j.inffus.2023.03.008

Bianchini, S., Müller, M., & Pelletier, P. (2020). Deep Learning in Science. *arXiv preprint arXiv:2009.01575*.

Bose, P., Srinivasan, S., Sleeman IV, W. C., Palta, J., Kapoor, R., & Ghosh, P. (2021). A survey on recently named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, *11*(18), 8319. https://doi.org/10.3390/app11188319

Braşoveanu, A. M., & Andonie, R. (2020, September). Visualizing transformers for nlp: a brief survey. In *2020 24th International Conference Information Visualisation (IV)* (pp. 270-279). IEEE. https://ieeexplore.ieee.org/abstract/document/9373074/

Beets, M. W., Weaver, R. G., Ioannidis, J. P., Geraci, M., Brazendale, K., Decker, L., Okely, A. D., Lubans, D., Van Sluijs, E., Jago, R., Turner-McGrievy, G., Thrasher, J.,

Li, X., & Milat, A. J. (2020). Identification and evaluation of risk of generalizability biases in pilot versus efficacy/effectiveness trials: A systematic review and meta-analysis. *International Journal of Behavioral Nutrition and Physical Activity*, *17*(1). https://doi.org/10.1186/s12966-020-0918-y

Bertsimas, D., & Öztürk, B. (2023). Global optimization via optimal decision trees. *Journal of Global Optimization*. https://doi.org/10.1007/s10898-023-01311-x

Biswas, Milon & Kaiser, M. Shamim & Al Mamun, Shamim & Hossain, Mohammad & Rahman, Muhammad. (2021). An XAI based Autism Detection: The Context Behind the Detection. 10.1007/978-3-030-86993-9_40.

Bacanin, N., Stoean, C., Zivkovic, M., Rakic, M., Strulak-Wójcikiewicz, R., & Stoean, R. (2023). On the benefits of using Metaheuristics in the Hyperparameter tuning of deep learning models for energy load forecasting. *Energies*, *16*(3), 1434. https://doi.org/10.3390/en16031434

Bafna, P. B., & R., J. (2020). An application of Zipf's law for prose and verse corpora neutrality for Hindi and Marathi languages. *International Journal of Advanced Computer Science and Applications*, *11*(3). https://doi.org/10.14569/ijacsa.2020.0110331

Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., ... & Sardanelli, F. (2021). AI applications to medical images: From machine learning to deep Learning. *Physica Medica*, *83*, 9-24.

Categorizing patient concerns using natural language processing techniques. *BMJ health & care informatics*, *28*(1). https://doi.org/10.1136%2Fbmjhci-2020-100274

Curtis, M. (2020). Toward understanding secondary teachers' decisions to adopt geospatial technologies: An examination of Everett Rogers' diffusion of innovation framework. *Journal of Geography*, *119*(5), 147-158.

Comparison and Analysis of SampleCNN Architectures for Audio Classification | IEEE Journals & Magazine | IEEE Xplore. (n.d.). Ieeexplore.ieee.org. Retrieved January 3, 2024, from https://ieeexplore.ieee.org/abstract/document/8681654

Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications. https://doi.org/10.1007/s00521-022-07049-z

Chaichulee, S., Promchai, C., Kaewkomon, T., Kongkamol, C., Ingviya, T., & Sangsupawanich, P. (2022). Multi-label classification of symptom terms from free-text bilingual adverse drug reaction reports using natural language processing. *PLOS ONE*, *17*(8), e0270595. https://doi.org/10.1371/journal.pone.0270595

Chandrapati, L. M., & Rao, C. K. (2023). Integrated assessment of teaching efficacy: A natural language processing approach. *International Journal of Advanced Computer Science and Applications*, *14*(1). https://doi.org/10.14569/ijacsa.2023.01401101

Chintalapudi, N., Angeloni, U., Battineni, G., Di Canio, M., Marotta, C., Rezza, G., Sagaro, G. G., Silenzi, A., & Amenta, F. (2022). LASSO regression modeling on prediction of medical terms among seafarers' health documents using tidy text mining. *Bioengineering*, *9*(3), 124. https://doi.org/10.3390/bioengineering9030124

Chintalapudi, N., Battineni, G., Canio, M. D., Sagaro, G. G., & Amenta, F. (2021). Text mining with sentiment analysis on seafarers' medical documents. *International Journal of Information Management Data Insights*, *1*(1), 100005. https://doi.org/10.1016/j.jjimei.2020.100005

Chatzinikolaou, T., Vogiatzi, E., Kousis, A., & Tjortjis, C. (2022). Smart healthcare support using data mining and machine learning. *IoT and WSN based Smart Cities: A Machine Learning Perspective*, 27-48. https://doi.org/10.1007/978-3-030-84182-9_3

Chen, P., Lin, C., & Wu, W. (2020). Big data management in healthcare: Adoption challenges and implications. *International Journal of Information Management, 53,* 102078. https://doi.org/10.1016/j.ijinfomgt.2020.102078

Cheng, C., Beauchamp, A., Elsworth, G. R., & Osborne, R. H. (2020). Applying the electronic health literacy lens: Systematic review of electronic health interventions targeted at socially disadvantaged groups. *Journal of Medical Internet Research*, *22*(8), e18476. https://doi.org/10.2196/18476

Cho, I., Lee, M., & Kim, Y. (2020). What are the main patient safety concerns of healthcare stakeholders: A mixed-method study of web-based text. *International Journal of Medical Informatics*, *140*, 104162. https://doi.org/10.1016/j.ijmedinf.2020.104162

Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., Bennett, C., Hawken, S., Magwood, O., Sheikh, Y., McInnes, M., & Holzinger, A. (2023). Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 329-341. https://doi.org/10.1109/tpami.2022.3145392

Casey, A., Davidson, E., Poon, M., Dong, H., Duma, D., Grivas, A., Grover, C., Suárez-Paniagua, V., Tobin, R., Whiteley, W., Wu, H., & Alex, B. (2021). A systematic review of natural language processing applied to radiology reports. *BMC Medical Informatics and Decision Making*, *21*(1). https://doi.org/10.1186/s12911-021-01533-7

Chen, R., Dewi, C., Huang, S., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00327-4

Chu, A., Squirrell, D., Phillips, A. M., & Vaghefi, E. (2020). Essentials of a robust deep learning system for diabetic retinopathy screening: A systematic literature review. *Journal of Ophthalmology*, *2020*, 1-11. https://doi.org/10.1155/2020/8841927

Das, S., Tariq, A., Santos, T., Sai Sandeep Kantareddy, & Banerjee, I. (2023). Recurrent Neural Networks (RNNs): Architectures, Training Tricks, and Introduction to Influential Research. Neuromethods, 117–138. https://doi.org/10.1007/978-1-0716-3195-9_4

Deshmukh, S. S. (2023, June). Progress in Machine Learning Techniques for Stock Market Movement Forecast. In *Proceedings of the International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022)* (Vol. 105, p. 69). Springer Nature.https://doi.org/10.2991/978-94-6463-136-4_9

Dobbins, N. J., Mullen, T., Uzuner, Ö., & Yetisgen, M. (2022). The Leaf Clinical Trials Corpus is a new resource for query generation from clinical trial eligibility criteria. *Scientific Data*, *9*(1), 490. https://doi.org/10.1038/s41597-022-01521-0

Dogra, V., Verma, S., Kavita, Chatterjee, P., Shafi, J., Choi, J., & Ijaz, M. F. (2022). A Complete Process of Text Classification System Using State-of-the-Art NLP Models. *Computational Intelligence and Neuroscience*, *2022*, 1–26. https://doi.org/10.1155/2022/1883698

Dey, L., Chakraborty, S., & Mukhopadhyay, A. (2020). Machine learning techniques for sequence-based prediction of viral–host interactions between SARS-Cov-2 and human proteins. *Biomedical Journal*, *43*(5), 438-450. https://doi.org/10.1016/j.bj.2020.08.003

Dishar, H. K., & Muhammed, L. A. (2023). A review of the Overfitting problem in convolution neural network and remedy approaches. *Journal of Al-Qadisiyah for Computer Science and Mathematics*, *15*(2). https://doi.org/10.29304/jqcm.2023.15.2.1240

Drake, C., Batchelder, H., Lian, T., Cannady, M., Weinberger, M., Eisenson, H., Esmaili, E., Lewinski, A., Zullig, L. L., Haley, A., Edelman, D., & Shea, C. M. (2021).

Implementation of social needs screening in primary care: A qualitative study using the health equity implementation framework. https://doi.org/10.21203/rs.3.rs-455846/v1

Dai, H., Su, C., Lee, Y., Zhang, Y., Wang, C., Kuo, C., & Wu, C. (2021). Deep learning-based natural language processing for screening psychiatric patients. *Frontiers in Psychiatry, 11*. https://doi.org/10.3389/fpsyt.2020.533949

Davoudi, A., Tissot, H., Doucette, A., Gabriel, P. E., Parikh, R., Mowery, D. L., & Miranda, S. (2021). Using natural language processing to classify serious illness communication with oncology patients. https://doi.org/10.1101/2021.08.20.21262082

Daviran, M., Maghsoudi, A., Ghezelbash, R., & Pradhan, B. (2021). A new strategy for spatial predictive mapping of mineral prospectivity: Automated hyperparameter tuning of random forest approach. *Computers & Geosciences, 148*, 104688. https://doi.org/10.1016/j.cageo.2021.104688

Dou, B., Zhu, Z., Merkurjev, E., Ke, L., Chen, L., Jiang, J., Zhu, Y., Liu, J., Zhang, B., & Wei, G. (2023). Machine learning methods for small data challenges in molecular science. *Chemical Reviews*, *123*(13), 8736-8780. https://doi.org/10.1021/acs.chemrev.3c00189

Dube, L., & Verster, T. (2023). Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models. *Data Science in Finance and Economics*, *3*(4), 354-379. https://doi.org/10.3934/dsfe.2023021

Efendy, B., Eko Mursito Budi, Estiyanti Ekawati, Mochamad Arief Soleh, & Herry Nugraha. (2023). Leakage prediction on superheater in boiler with hierarchical clustering and Naïve Bayes classification. AIP Conference Proceedings. https://doi.org/10.1063/5.0124509

Emadi, M., Delavari, S., & Bayati, M. (2021). Global socioeconomic inequality in the burden of communicable and non-communicable diseases and injuries: An analysis on global burden of disease study 2019. *BMC Public Health*, *21*(1). https://doi.org/10.1186/s12889-021-11793-7

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Banyatsang, M., & Tabona, O. (2021). A survey on missing data in machine learning. https://doi.org/10.21203/rs.3.rs-535520/v1

Ennab, M., & Mcheick, H. (2022). Survey of COVID-19 prediction models and their limitations. *International Journal of Intelligent Information Systems*, *11*(2), 14. https://doi.org/10.11648/j.ijiis.20221102.11

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis. *Informatics*, *8*(4), 79. https://doi.org/10.3390/informatics8040079

Elkin, P. L., Mullin, S., Mardekian, J., Crowner, C., Sakilay, S., Sinha, S., Brady, G., Wright, M., Nolen, K., Trainer, J., Koppel, R., Schlegel, D., Kaushik, S., Zhao, J., Song, B., & Anand, E. (2021). Using artificial intelligence with natural language processing to combine electronic health record's structured and free text data to identify nonvalvular atrial fibrillation to decrease strokes and death: Evaluation and case-control study. *Journal of Medical Internet Research*, *23*(11), e28946. https://doi.org/10.2196/28946

Erickson, B. J., & Kitamura, F. (2021). Magician's corner: 8: How to connect an artificial intelligence tool to PACS. *Radiology: Artificial Intelligence*, *3*(1), e200105. https://doi.org/10.1148/ryai.2021200105

Fagherazzi, G., Fischer, A., Ismael, M., & Despotovic, V. (2021). Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. Digital Biomarkers, 5(1), 78–88. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8138221/

Fairie, P., Zhang, Z., D'Souza, A. G., Walsh, T., Quan, H., & Santana, M. J. (2021).

Fahy, N., Greenhalgh, T., & Shaw, S. (2020). PHOENIX: A new framework for applying psychological theories to the adoption of innovations by healthcare professionals.

Faieq, A. K., & Mijwil, M. M. (2022). Prediction of heart diseases utilizing support vector machine and artificial neural network. *Indonesian Journal of Electrical Engineering and Computer Science*, *26*(1), 374. https://doi.org/10.11591/ijeecs.v26.i1.pp374-380

Fang, C., Markuzon, N., Patel, N., & Rueda, J. (2022). Natural language processing for automated classification of qualitative data from interviews of patients with cancer. *Value in Health*, *25*(12), 1995-2002. https://doi.org/10.1016/j.jval.2022.06.004

Faris, H., Habib, M., Faris, M., Elayan, H., & Alomari, A. (2021). An intelligent multimodal medical diagnosis system based on patients' medical questions and structured symptoms for telemedicine. *Informatics in Medicine Unlocked*, *23*, 100513. https://doi.org/10.1016/j.imu.2021.100513

Fernandes, M., Sun, H., Jain, A., Alabsi, H. S., Brenner, L. N., Ye, E., Ge, W., Collens, S. I., Leone, M. J., Das, S., Robbins, G. K., Mukerji, S. S., & Westover, M. B. (2021). Classification of the disposition of patients hospitalized with COVID-19: Reading discharge summaries using natural language processing. *JMIR Medical Informatics*, *9*(2), e25457. https://doi.org/10.2196/25457

FLEISCHER, Y., BIEHLER, R., & SCHULTE, C. (2022). Teaching and learning data-driven machine learning with educationally designed jupyter notebooks. *STATISTICS*

EDUCATION RESEARCH JOURNAL, *21*(2), 7. https://doi.org/10.52041/serj.v21i2.61

Falissard, B. (2021). The future of evaluation of child and adolescent psychiatric treatments. *IACAPAP ArXiv*. https://doi.org/10.14744/iacapaparxiv.2020.20007

Flemotomos, N., Martinez, V. R., Chen, Z., Singla, K., Ardulov, V., Peri, R., Caperton, D. D., Gibson, J., Tanana, M. J., Georgiou, P., Van Epps, J., Lord, S. P., Hirsch, T., Imel, Z. E., Atkins, D. C., & Narayanan, S. (2021). Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods*, *54*(2), 690-711. https://doi.org/10.3758/s13428-021-01623-4

GOYAL, A. A. (2023). The Role of Machine Learning in Natural Language Processing and Computer Vision.

Ge, C., Susilo, W., Baek, J., Liu, Z., Xia, J., & Fang, L. (2022). Revocable attribute-based encryption with data integrity in clouds. *IEEE Transactions on Dependable and Secure Computing*, *19*(5), 2864-2872. https://doi.org/10.1109/tdsc.2021.3065999

Georges, P., & Seckin, A. (2022). Music information visualization and classical composers discovery: An application of network graphs, multidimensional scaling, and support vector machines. *Scientometrics*, *127*(5), 2277-2311. https://doi.org/10.1007/s11192-022-04331-8

Ghosh, S., Das, N., & Nasipuri, M. (2019). Reshaping inputs for convolutional neural network: Some common and uncommon methods. *Pattern Recognition*, *93*, 79–94. https://doi.org/10.1016/j.patcog.2019.04.009

Golinelli, D., Boetto, E., Carullo, G., Nuzzolese, A. G., Landini, M. P., & Fantini, M. P. (2020). Adoption of digital technologies in health care during the COVID-19 pandemic: Systematic review of early scientific literature. *Journal of Medical Internet Research*, *22*(11), e22280. https://doi.org/10.2196/22280

Golinelli, D., Boetto, E., Carullo, G., Nuzzolese, A. G., Landini, M. P., & Fantini, M. P. (2020). Adoption of digital technologies in health care during the COVID-19 pandemic: Systematic review of early scientific literature. *Journal of Medical Internet Research*, *22*(11), e22280. https://doi.org/10.2196/22280

Gonon, L., Grigoryeva, L., & Ortega, J. (2023). Approximation bounds for random neural networks and reservoir systems. *The Annals of Applied Probability*, *33*(1). https://doi.org/10.1214/22-aap1806

Görtler, J., Hohman, F., Moritz, D., Wongsuphasawat, K., Ren, D., Nair, R., Kirchner, M., & Patel, K. (2022). Neo: Generalizing confusion matrix visualization to hierarchical and multi-output labels. *CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3491102.3501823

Gürsakal, N., Çelik, S., & Özdemir, S. (2022). High-frequency words have higher frequencies in Turkish social sciences article. *Quality & Quantity*, *57*(2), 1865-1887. https://doi.org/10.1007/s11135-022-01444-3

Heys, M., Kesler, E., Sassoon, Y., Wilson, E., Fitzgerald, F., Gannon, H., Hull-Bailey, T., Chimhini, G., Khan, N., Cortina-Borja, M., Nkhoma, D., Chiyaka, T., Stevenson, A., Crehan, C., Chiume, M. E., & Chimhuya, S. (2022). Development and implementation experience of a learning healthcare system for facility-based newborn care in low resource settings: The Neotree. Learning Health Systems, 7(1). https://doi.org/10.1002/lrh2.10310

Hisamitsu, T., Oikawa, M., & Kido, K. (2016). Care cycle optimization using digital solutions. *Hitachi Review*, *65*(9), 399. https://www.hitachi.com/rev/archive/2016/r2016_09/pdf/r2016_09_105.pdf

Hasan, M. I., Ali, M. S., Rahman, M. H., & Islam, M. K. (2022). Automated detection and characterization of colon cancer with deep Convolutional neural networks. *Journal of Healthcare Engineering, 2022,* 1-12. https://doi.org/10.1155/2022/5269913

Haulcy, R. M., & Glass, J. (2021). Classifying Alzheimer's disease using audio and text-based representations of speech. *Frontiers in Psychology*, *11*, 624137. https://d1wqtxts1xzle7.cloudfront.net/94942161/ASTESJ_030437-libre.pdf?1669610765=&response-content-disposition=inline%3B+filename%3DAmplitude_Frequency_Analysis_of_Emotiona.pdf&Expires=1712703661&Signature=M-tmk2vGHdr5xZiMJ3EJEeoY5gQ~upPPzO0i-XTfgJhRYGaffxUb12wmjEffM78AWU7EFTn1dQX9uxRoLi~8CRVyJnhy4f-Emgzrlt4yVyi9J2r7VTWwFZEqCELCBXWnE3PRryBkzc1RWaLlJQFMWb06dTcnaCRsyFVkT4hjgu6TUFUbcuuAA6iYT5wtxVVKKvEFK-O3kyWvH8j3wGUukprcv9Jci~zAAYu-vy29BAIfNnRUPeqfEC2DBsrRNPiOyB6Rh3z-W8C1BRBy9DfiTvuzAGiq9zsAlUDdRO~CrPrsYJ5XKqStFUblJeChzcY5mF07C0fK~xu5oz6~7MMZw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Hussain, Shahid & Irshad, Reyazur & Hussain, Ihtisham & Alattab, Ahmed & Yousif, Adil & Alsaiari, Omar & Ibrahim, Elshareef. (2023). A Novel Artificial Spider Monkey Based Random Forest Hybrid Framework for Monitoring and Predictive Diagnoses of Patients Healthcare. IEEE Access. PP. 10.1109/ACCESS.2023.3297957.

Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using Ensembling of different machine learning classifiers. *IEEE Access*, *8*, 76516-76531. https://doi.org/10.1109/access.2020.2989857

Hasnain, M., Pasha, M. F., Ghani, I., Imran, M., Alzahrani, M. Y., & Budiarto, R. (2020). Evaluating trust prediction and confusion matrix measures for web services ranking. *IEEE Access, 8*, 90847-90861. https://doi.org/10.1109/access.2020.2994222

Hasnain, M., Pasha, M. F., Ghani, I., Imran, M., Alzahrani, M. Y., & Budiarto, R. (2020). Evaluating trust prediction and confusion matrix measures for web services ranking. *IEEE Access, 8*, 90847-90861. https://doi.org/10.1109/access.2020.2994222

Heydarian, M., Doyle, T. E., & Samavi, R. (2022). MLCM: Multi-label confusion matrix. *IEEE Access, 10*, 19083-19095. https://doi.org/10.1109/access.2022.3151048

Holzinger, A., Dehmer, M., Emmert-Streib, F., Cucchiara, R., Augenstein, I., Ser, J. D., Samek, W., Jurisica, I., & Díaz-Rodríguez, N. (2022). undefined. *Information Fusion*, *79*, 263-278. https://doi.org/10.1016/j.inffus.2021.10.007

Horigome, T., Hino, K., Toyoshiba, H., Shindo, N., Funaki, K., Eguchi, Y., Kitazawa, M., Fujita, T., Mimura, M., & Kishimoto, T. (2022). Identifying neurocognitive disorder using vector representation of free conversation. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-16204-4

https://www.kaggle.com/datasets/paultimothymooney/medical-speech-transcription-and-intent)

Jayakumar, S., Sounderajah, V., Normahani, P., Harling, L., Markar, S. R., Ashrafian, H., & Darzi, A. (2022). Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. *Npj Digital Medicine*, *5*(1). https://doi.org/10.1038/s41746-021-00544-y

Iroju, O. G., & Olaleke, J. O. (2015). A systematic review of natural language processing in healthcare. *International Journal of Information Technology and Computer Science*, *7*(8), 44-50. https://doi.org/10.5815/ijitcs.2015.08.07

Iwana, B. K., & Uchida, S. (2021). An empirical survey of data augmentation for time series classification with neural networks. *PLOS ONE, 16*(7), e0254841. https://doi.org/10.1371/journal.pone.0254841

Inoue, T., Endo, T., Suzuki, S., Uenohara, H., & Tominaga, T. (2019). Multivariate analysis of acute magnetic resonance imaging predicts neurological improvements in patients with cervical spinal cord injury after early surgical decompression. *Neurosurgery, 66*(Supplement 1), 310-841. https://doi.org/10.1093/neuros/nyz310_841

Jain, S., Naicker, D., Raj, R., Patel, V., Hu, Y.-C., Srinivasan, K., & Jen, C.-P. (2023). Computational Intelligence in Cancer Diagnostics: A Contemporary Review of Smart Phone Apps, Current Problems, and Future Research Potentials. Diagnostics, 13(9), 1563. https://doi.org/10.3390/diagnostics13091563

Johri, P., Khatri, S. K., Al-Taani, A. T., Sabharwal, M., Suvanov, S., & Kumar, A. (2021). Natural language processing: History, evolution, application, and future work. In *Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020* (pp. 365-375). Springer Singapore.https://doi.org/10.1007/978-981-15-9712-1_31

Jalilov, S., Chen, Y., Quang, N. H., Nguyen, M. N., Leighton, B., Paget, M., & Lazarow, N. (2021). Estimation of urban land-use efficiency for sustainable development by integrating over 30-Year Landsat imagery with population data: A case study of ha long, Vietnam. *Sustainability, 13*(16), 8848. https://doi.org/10.3390/su13168848

Jeong, H., Wang, H., & Calmon, F. P. (2022). Fairness without imputation: A decision tree approach for fair prediction with missing values. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(9), 9558-9566. https://doi.org/10.1609/aaai.v36i9.21189

Jones, R., Manly, J., Langa, K., Ryan, L., Levine, D., McCammon, R., & Weir, D. (2020). Factor structure of the harmonized cognitive assessment protocol neuropsychological battery in the health and retirement study. https://doi.org/10.31234/osf.io/rvmhj

Josephson, A., & Smale, M. (2020). What do you mean by “Informed consent”? Ethics in economic development Research†. *Applied Economic Perspectives and Policy*, *43*(4), 1305-1329. https://doi.org/10.1002/aepp.13112

Jaskowiak, P. A., Costa, I. G., & Campello, R. J. (2022). The area under the ROC curve as a measure of clustering quality. *Data Mining and Knowledge Discovery*, *36*(3), 1219-1245. https://doi.org/10.1007/s10618-022-00829-0

Joseph, J., Moore, Z. E., Patton, D., O'Connor, T., & Nugent, L. E. (2020). undefined. *Journal of Clinical Nursing*, *29*(13-14), 2125-2137. https://doi.org/10.1111/jocn.15261

Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, *7*(2), 139-172. https://www.tandfonline.com/doi/abs/10.1080/23270012.2020.1756939

Kasongo, S. M. (2022). A deep learning technique for intrusion detection system using a Recurrent Neural Networks based framework. Computer Communications. https://doi.org/10.1016/j.comcom.2022.12.010

Kobritz, M., Patel, V., Rindskopf, D., Demyan, L., Jarrett, M., Coppa, G., & Antonacci, A. C. (2023). Practice-Based Learning and Improvement: Improving Morbidity and Mortality Review Using Natural Language Processing. Journal of Surgical Research, 283, 351–356. https://doi.org/10.1016/j.jss.2022.10.075

Kozłowska, U., & Sikorski, T. (2021). The Implementation of the Soviet Healthcare Model in 'People's Democracy'Countries—the Case of Post-war Poland (1944–1953). *Social History of Medicine*, *34*(4), 1185-1211.

Kruse, C. S., Goswamy, R., Raval, Y., & Marawi, S. (2016). Challenges and opportunities of big data in health care: A systematic review. *JMIR Medical Informatics*, *4*(4), e38. https://doi.org/10.2196/medinform.5359

Kasthuri, E., & Balaji, S. (2023). Natural language processing and deep learning chatbot using long short term memory algorithm. *Materials Today: Proceedings*, *81*, 690-693. https://doi.org/10.1016/j.matpr.2021.04.154

Kaur, J., Kaur, J., Kapoor, S., & Singh, H. (2021). Design & development of customizable web API for interoperability of antimicrobial resistance data. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-90601-z

Khatoon, A. (2020). A blockchain-based smart contract system for healthcare management. *Electronics*, *9*(1), 94. https://doi.org/10.3390/electronics9010094

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). 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, J. W., & Lee, E. J. (2021). Effect of General Hospital nurses' perception of patient safety culture and organizational communication satisfaction on safe care. *Forum of Public Safety and Culture*, *11*, 131-143. https://doi.org/10.52902/kjsc.2021.11.131

Kostova, D., Richter, P., Van Vliet, G., Mahar, M., & Moolenaar, R. L. (2021). The role of noncommunicable diseases in the pursuit of global health security. *Health Security, 19*(3), 288-301. https://doi.org/10.1089/hs.2020.0121

Križanić, S. (2020). Educational data mining using cluster analysis and decision tree technique: A case study. *International Journal of Engineering Business Management*, *12*, 184797902090867. https://doi.org/10.1177/1847979020908675

Kurani, A., Doshi, P., Vakharia, A., & Shah, M. (2021). A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. *Annals of Data Science*, *10*(1), 183-208. https://doi.org/10.1007/s40745-021-00344-x

Kaswan, K. S., Gaur, L., Dhatterwal, J. S., & Kumar, R. (2021). AI-based natural language processing for the generation of meaningful information electronic health record (EHR) data. *Advanced AI Techniques and Applications in Bioinformatics*, 41-86. https://doi.org/10.1201/9781003126164-3

Kerexeta, J., Torres, J., Muro, N., Rebescher, K., & Larburu, N. (2020). Adaptative clinical decision support system using machine learning and authoring tools. *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies*, 95-105. https://doi.org/10.5220/0008952200002513

Kernbach, J. M., & Staartjes, V. E. (2021). Foundations of machine learning-based clinical prediction modeling: Part II—Generalization and Overfitting. *Acta Neurochirurgica Supplement*, 15-21. https://doi.org/10.1007/978-3-030-85292-4_3

Kieu, S. T., Bade, A., Hijazi, M. H., & Kolivand, H. (2020). A survey of deep learning for lung disease detection on medical images: State-of-the-Art, taxonomy, issues and future directions. *Journal of Imaging*, *6*(12), 131. https://doi.org/10.3390/jimaging6120131

Kowsari, K., Sali, R., Ehsan, L., Adorno, W., Ali, A., Moore, S., Amadi, B., Kelly, P., Syed, S., & Brown, D. (2020). HMIC: Hierarchical medical image classification, a

deep learning approach. *Information*, *11*(6), 318. https://doi.org/10.3390/info11060318

Krstinić, D., Braović, M., Šerić, L., & Božić-Štulić, D. (2020). Multi-label classifier performance evaluation with confusion matrix. *Computer Science & Information Technology*. https://doi.org/10.5121/csit.2020.100801

Leichter, H. (1979). A Comparative Approach to Policy Analysis Health Care Policy in Four Nations. Cambridge University Press.

Li, I., Pan, J., Goldwasser, J., Verma, N., Wong, W. P., Nuzumlalı, M. Y., Rosand, B., Li, Y., Zhang, M., Chang, D., Taylor, R. A., Krumholz, H. M., & Radev, D. (2022). Neural natural language processing for unstructured data in electronic health records: A review. *Computer Science Review*, *46*, 100511. https://doi.org/10.1016/j.cosrev.2022.100511

Langa, K., Ryan, L., McCammon, R., Jones, R., Manly, J., Levine, D., Sonnega, A., Farron, M., & Weir, D. (2019). The health and retirement study harmonized cognitive assessment protocol project: Study design and methods. *Neuroepidemiology*, *54*(1), 64-74. https://doi.org/10.1159/000503004

Lavanya, P., & Sasikala, E. (2021). Deep learning techniques on text classification using natural language processing (NLP) in social healthcare network: A comprehensive survey. *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*. https://doi.org/10.1109/icspc51351.2021.9451752

Leary, A., Bushe, D., Oldman, C., Lawler, J., & Punshon, G. (2021). A thematic analysis of the prevention of future deaths reports in healthcare from HM coroners in England and Wales 2016–2019. *Journal of Patient Safety and Risk Management*, *26*(1), 14-21. https://doi.org/10.1177/2516043521992651

Le Glaz, A., Haralambous, Y., Kim-Dufor, D., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., DeVylder, J., Walter, M., Berrouiguet, S., & Lemey, C. (2021). Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research*, *23*(5), e15708. https://doi.org/10.2196/15708

Lezhenin, Iurii & Bogach, Natalia & Pyshkin, Evgeny. (2019). Urban Sound Classification using Long Short-Term Memory Neural Network. 57-60. 10.15439/2019F185.

Liang, Y., Liang, W., & Jia, J. (2023). Structural vibration signal Denoising using stacking ensemble of hybrid CNN-RNN. *Advances in Artificial Intelligence and Machine Learning*, *03*(02), 1110-1122. https://doi.org/10.54364/aaiml.2023.1165

Liu, B. J., & Huang, H. (2020). Picture archiving and communication systems and electronic medical records for the healthcare enterprise. *Biomedical Information Technology*, 105-164. https://doi.org/10.1016/b978-0-12-816034-3.00004-3

Lu, Z., Sim, J., Wang, J. X., Forrest, C. B., Krull, K. R., Srivastava, D., Hudson, M. M., Robison, L. L., Baker, J. N., & Huang, I. (2021). Natural language processing and machine learning methods to characterize unstructured patient-reported outcomes: Validation study. *Journal of Medical Internet Research*, *23*(11), e26777. https://doi.org/10.2196/26777

Luo, X., Gandhi, P., Storey, S., & Huang, K. (2022). A deep language model for symptom extraction from clinical text and its application to extract COVID-19 symptoms from social media. *IEEE Journal of Biomedical and Health Informatics*, *26*(4), 1737-1748. https://doi.org/10.1109/jbhi.2021.3123192

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, *37*(3), 50-60. https://doi.org/10.1109/msp.2020.2975749

Lilhore, U. K., Poongodi, M., Kaur, A., Simaiya, S., Algarni, A. D., Elmannai, H., Vijayakumar, V., Tunze, G. B., & Hamdi, M. (2022). Hybrid model for detection of cervical cancer using causal analysis and machine learning techniques. *Computational and Mathematical Methods in Medicine*, *2022*, 1-17. https://doi.org/10.1155/2022/4688327

MA, Jun & Ding, Yuexiong & Gan, Vincent & Lin, Changqing & WAN, Zhiwei. (2019). Spatiotemporal Prediction of PM2.5 Concentrations at Different Time Granularities Using IDW-BLSTM. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2932445.

Malik, H., Bashir, U., & Ahmad, A. (2022). Multi-classification neural network model for detection of abnormal heartbeat audio signals. Biomedical Engineering Advances, 4, 100048. https://doi.org/10.1016/j.bea.2022.100048

Mishra, S. B., & Alok, S. (2022). Handbook of research methodology. https://www.researchgate.net/publication/319207471_HANDBOOK_OF_RESEARCH_METHODOLOGY?enrichId=rgreq-6be5390a6f24699c882b5c3de1cc9f78-XXX&enrichSource=Y292ZXJQYWdlOzMxOTIwNzQ3MTtBUzo3MTQ4NTgxNDAwMTY2NDJAMTU0NzQ0Njg3MDk1Mg%3D%3D&el=1_x_2&_esc=publicationCoverPdf

Moez Krichen. (2023). Convolutional Neural Networks: A Survey. Computers, 12(8), 151–151. https://doi.org/10.3390/computers12080151

Mohammadi, M. M., Poursaberi, R., & Salahshoor, M. R. (2018). Evaluating the adoption of evidence-based practice using Rogers's diffusion of innovation theory: A model testing study. Health Promotion Perspectives, 8(1), 25–32. https://doi.org/10.15171/hpp.2018.03

Moorhead, L. (2021, June 17). *Resize multiple images to be the same size.* Miro. https://community.miro.com/ask-the-community-45/resize-multiple-images-to-be-the-same-size-5101

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1), 91-99. https://doi.org/10.1016/j.gltp.2022.04.020

Mahdi, H. (2024, November 1). Dissertation-Manuscript. GitHub https://github.com/HAMEEMM/Dissertation-Manuscript

Malgaroli, M., Hull, T. D., Zech, J. M., & Althoff, T. (2023). Natural language processing for mental health interventions: A systematic review and research framework. *Translational Psychiatry*, *13*(1). https://doi.org/10.1038/s41398-023-02592-2

Malone, M., Ferguson, P., Rogers, A., Mackenzie, I. S., Rorie, D. A., & MacDonald, T. M. (2021). When innovation outpaces regulations: The legal challenges for direct-to-patient supply of investigational medicinal products. *British Journal of Clinical Pharmacology*, *88*(3), 1115-1142. https://doi.org/10.1111/bcp.15040

Marra, M. (2018). Astrophysicists and physicists as creators of arXiv-based commenting resources for their research communities. An initial survey. *Information Services & Use*, *37*(4), 371-387. https://doi.org/10.3233/isu-170856

Martinez, R., Lloyd-Sherlock, P., Soliz, P., Ebrahim, S., Vega, E., Ordunez, P., & McKee, M. (2020). Trends in premature avertable mortality from non-communicable diseases for 195 countries and territories, 1990–2017: A population-based study. *The Lancet Global Health*, *8*(4), e511-e523. https://doi.org/10.1016/s2214-109x(20)30035-8

Martínez Munoz, G., Chikh, M. A., Settouti, N., & Guilal, R. (2022). Feature importance analysis for a highly unbalanced multiple myeloma data classification. *International*

Journal of Medical Engineering and Informatics, *1*(1), 1. https://doi.org/10.1504/ijmei.2022.10046878

Mehedi Hassan, M., Mollick, S., & Yasmin, F. (2022). An unsupervised cluster-based feature grouping model for early diabetes detection. *Healthcare Analytics*, *2*, 100112. https://doi.org/10.1016/j.health.2022.100112

Meshram, S. G., Singh, V. P., Kisi, O., Karimi, V., & Meshram, C. (2020). Application of artificial neural networks, support vector machine and multiple Model-ANN to sediment yield prediction. *Water Resources Management*, *34*(15), 4561-4575. https://doi.org/10.1007/s11269-020-02672-8

Mirbabaie, M., Stieglitz, S., & Frick, N. R. (2021). Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. *Health and Technology*, *11*(4), 693-731. https://doi.org/10.1007/s12553-021-00555-5

Mitchell, F., Nørreklit, H., Nørreklit, L., Cinquini, L., Koeppe, F., Magnacca, F., Mauro, S. G., Jakobsen, M., Korhonen, T., Laine, T., & Liboriussen, J. M. (2021). Evaluating performance management of COVID-19 reality in three European countries: A pragmatic constructivist study. *Accounting, Auditing & Accountability Journal*, *34*(6), 1345-1361. https://doi.org/10.1108/aaaj-08-2020-4778

Montes, D., Pongpatapee Peerapatanapokin, Schultz, J., Guo, C., Jiang, W., & Davis, J. C. (2022). Discrepancies among pre-trained deep neural networks: a new threat to model zoo reliability. Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. https://doi.org/10.1145/3540250.3560881

Mou, G., Li, Y., & Lee, K. (2021). Reducing and exploiting data augmentation noise through Meta Reweighting contrastive learning for text classification. *2021 IEEE*

International Conference on Big Data (Big Data). https://doi.org/10.1109/bigdata52589.2021.9671510

Mozersky, J., Walsh, H., Parsons, M., McIntosh, T., Baldwin, K., & DuBois, J. M. (2020). Are we ready to share qualitative research data? Knowledge and preparedness among qualitative researchers, IRB members, and data repository curators. *IASSIST Quarterly*, *43*(4), 1-23. https://doi.org/10.29173/iq952

Mao, C., Zhu, Q., Chen, R., & Su, W. (2023). Automatic medical specialty classification based on patients' description of their symptoms. *BMC Medical Informatics and Decision Making*, *23*(1). https://doi.org/10.1186/s12911-023-02105-7

Mukherjee, S. S., Yu, J., Won, Y., McClay, M. J., Wang, L., Rush, A. J., & Sarkar, J. (2020). Natural language processing-based Quantication of the mental state of psychiatric patients. *Computational Psychiatry*, *4*(0), 76. https://doi.org/10.1162/cpsy_a_00030

Mukhiya, S. K., Ahmed, U., Rabbi, F., Pun, K. I., & Lamo, Y. (2020). Adaptation of IDPT system based on patient-authored text data using NLP. *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. https://doi.org/10.1109/cbms49503.2020.00050

Murbach, M., Gerwe, B., Dawson-Elli, N., & Tsui, L. (2020). Impedance.py: A python package for electrochemical impedance analysis. *Journal of Open Source Software*, *5*(52), 2349. https://doi.org/10.21105/joss.02349

Malgaroli, M., Hull, T. D., Zech, J. M., & Althoff, T. (2023). Natural language processing for mental health interventions: A systematic review and research framework. *Translational Psychiatry*, *13*(1). https://doi.org/10.1038/s41398-023-02592-2

Malgaroli, M., Hull, T. D., Zech, J. M., & Althoff, T. (2023). Natural language processing for mental health interventions: A systematic review and research

framework. *Translational Psychiatry*, *13*(1). https://doi.org/10.1038/s41398-023-02592-2

Manibardo, E. L., Lana, I., & Del Ser, J. (2020). Transfer learning and online learning for traffic forecasting under different data availability conditions: Alternatives and pitfalls. *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. https://doi.org/10.1109/itsc45102.2020.9294557

Martinc, M., Haider, F., Pollak, S., & Luz, S. (2021). Temporal integration of text transcripts and acoustic features for Alzheimer's diagnosis based on spontaneous speech. *Frontiers in Aging Neuroscience*, *13*. https://doi.org/10.3389/fnagi.2021.642647

Merchant, H. A., Kow, C. S., & Hasan, S. S. (2021). COVID-19 first anniversary review of cases, hospitalization, and mortality in the UK. *Expert Review of Respiratory Medicine*, *15*(8), 973-978. https://doi.org/10.1080/17476348.2021.1890035

Morales-Sánchez, R., Montalvo, S., Riaño, A., Martínez, R., & Velasco, M. (2024). Early diagnosis of HIV cases by means of text mining and machine learning models on clinical notes. *Computers in Biology and Medicine*, *179*, 108830. https://doi.org/10.1016/j.compbiomed.2024.108830

Mouliou, D. S., Pantazopoulos, I., & Gourgoulianis, K. I. (2022). COVID-19 smart diagnosis in the emergency department: All-in in practice. *Expert Review of Respiratory Medicine*, *16*(3), 263-272. https://doi.org/10.1080/17476348.2022.2049760

Movahedi, F., Padman, R., & Antaki, J. F. (2023). Limitations of receiver operating characteristic curve on imbalanced data: Assist device mortality risk scores. *The Journal of Thoracic and Cardiovascular Surgery*, *165*(4), 1433-1442.e2. https://doi.org/10.1016/j.jtcvs.2021.07.041

Murtaza, G., Shuib, L., Abdul Wahab, A. W., Mujtaba, G., Mujtaba, G., Nweke, H. F., Al-garadi, M. A., Zulfiqar, F., Raza, G., & Azmi, N. A. (2019). Deep learning-based breast cancer classification through medical imaging modalities: State of the art and research challenges. *Artificial Intelligence Review*, *53*(3), 1655-1720. https://doi.org/10.1007/s10462-019-09716-5

Nawab, K., Ramsey, G., & Schreiber, R. (2020). Natural language processing to extract meaningful information from patient experience feedback. *Applied Clinical Informatics*, *11*(02), 242-252. 10.1055/s-0040-1708049

Newman-Toker, D. E., Wang, Z., Zhu, Y., Nassery, N., Saber Tehrani, A. S., Schaffer, A. C., Yu-Moe, C. W., Clemens, G. D., Fanai, M., & Siegal, D. (2020). Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: Toward a national incidence estimate using the "Big three". *Diagnosis*, *8*(1), 67-84. https://doi.org/10.1515/dx-2019-0104

Nijhawan, N. A., & Al-Shamsi, H. O. (2021). Palliative care in the United Arab Emirates (UAE). *Handbook of Healthcare in the Arab World*, 1-18. https://doi.org/10.1007/978-3-319-74365-3_102-1

Novelli, G., Biancolella, M., Latini, A., Spallone, A., Borgiani, P., & Papaluca, M. (2020). Precision medicine in non-communicable diseases. *High-Throughput*, *9*(1), 3. https://doi.org/10.3390/ht9010003

Nozari, H., Tavakkoli-Moghaddam, R., Ghahremani-Nahr, J., & Najafi, E. (2023, January 1). Chapter 8 - A conceptual framework for Artificial Intelligence of Medical Things (AIoMT) (Y. Maleh, A. A. A. El-Latif, K. Curran, P. Siarry, N. Dey, A. Ashour, & S. J. Fong, Eds.). ScienceDirect; Academic Press. https://www.sciencedirect.com/science/article/abs/pii/B9780323994217000076

Ngan, T., Haihua, C., Janet, J., Jay, B., & Junhua, D. (2021). Effect of class imbalance on the performance of machine learning-based network intrusion detection. *International Journal of Performability Engineering*, *17*(9), 741. https://doi.org/10.23940/ijpe.21.09.p1.741755

O'Cathain, A., Connell, J., Long, J., & Coster, J. (2020). 'Clinically unnecessary of emergency and urgent care: A realist review of patients' decision making. *Health Expectations*, *23*(1), 19-40.

Okwuashi, O., Ndehedehe, C. E., Olayinka, D. N., Eyoh, A., & Attai, H. (2021). Deep support vector machine for PolSAR image classification. *International Journal of Remote Sensing*, *42*(17), 6498-6536. https://doi.org/10.1080/01431161.2021.1939910

Ozaslan, A. (2020). Evaluation of institutionally reared children and adolescents in terms of mental health in Ankara. *IACAPAP ArXiv*. https://doi.org/10.14744/iacapaparxiv.2020.20001

Odhiambo Omuya, E., Onyango Okeyo, G., & Waema Kimwele, M. (2021). Feature selection for classification using principal component analysis and information gain. *Expert Systems with Applications*, *174*, 114765. https://doi.org/10.1016/j.eswa.2021.114765

Palanisamy, K., Singhania, D., & Yao, A. (2020). Rethinking CNN Models for Audio Classification. ArXiv:2007.11154 [Cs, Eess]. https://arxiv.org/abs/2007.11154

Pearson, K., League, R., Kent, M. L., McDevitt, R. C., Fuller, M., Jiang, R., Melton, S., Krishnamoorthy, V., Tetsu Ohnuma, Bartz, R. R., Cobert, J., & Raghunathan, K. (2023). Rogers' diffusion theory of innovation applied to the adoption of sugammadex in a nationwide sample of US hospitals. British Journal of Anaesthesia, 131(4), e114–e117. https://doi.org/10.1016/j.bja.2023.06.061

Pandey, D. K., Hunjra, A. I., Bhaskar, R., & Al-Faryan, M. A. (2023). Artificial intelligence, machine learning and big data in natural resources management: A comprehensive bibliometric review of literature spanning 1975–2022. *Resources Policy, 86*, 104250. https://doi.org/10.1016/j.resourpol.2023.104250

Pereira, G. B., Santos, E. A., & Maceno, M. M. (2020). Correction to: Process mining project methodology in healthcare: a case study in a tertiary hospital. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *9*(1). https://doi.org/10.1007/s13721-020-00247-6

Pajankar, A. (2020). Getting started with pandas. *Practical Python Data Visualization*, 117-136. https://doi.org/10.1007/978-1-4842-6455-3_8

Poulsen, M. N., Freda, P. J., Troiani, V., Davoudi, A., & Mowery, D. L. (2022). Classifying characteristics of opioid use disorder from hospital discharge summaries using natural language processing. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.850619

Passos, D., & Mishra, P. (2022). A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks. *Chemometrics and Intelligent Laboratory Systems*, *223*, 104520. https://doi.org/10.1016/j.chemolab.2022.104520

Qi, X., Zhou, Q., Dong, J., & Bao, W. (2023). Noninvasive automatic detection of Alzheimer's disease from spontaneous speech: A review. *Frontiers in Aging Neuroscience*, *15*. https://doi.org/10.3389/fnagi.2023.1224723

Qin, X., Liu, J., Wang, Y., Liu, Y., Deng, K., Ma, Y., Zou, K., Li, L., & Sun, X. (2021). Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews. *Journal of Clinical Epidemiology*, *133*, 121-129. https://doi.org/10.1016/j.jclinepi.2021.01.010

Razgan, M. A., Alrowily, A., Matham, R. N. A., Alghamdi, K. M., Shaabi, M., & Alssum, L. (2021). Using diffusion of innovation theory and sentiment analysis to analyze attitudes toward driving adoption by Saudi women. Technology in Society, 65, 1–11. Sciencedirect. https://doi.org/10.1016/j.techsoc.2021.101558

Ritter, E. (2021). Your Voice Gave You Away: The Privacy Risks of Voice-Inferred Information. *Duke LJ*, *71*, 735.

Rusk, N. (2016). Deep Learning. Nature Methods, 13(1), 35–35. https://doi.org/10.1038/nmeth.3707

Rahman Chowdhury, M., Ahmed, I., Sadeque, F., & Nur Yanhaona, M. (2023). Topic modeling using community detection on a word association graph. *Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings*. https://doi.org/10.26615/978-954-452-092-2_098

Radhika, A., Dengri, C., Kumar, A., & Singh, S. (2021). A structured maternity case record to improve quality of documentation in a tertiary care hospital, Delhi. *JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH*. https://doi.org/10.7860/jcdr/2021/45693.14647

Rahim, T., Hassan, S. A., & Shin, S. Y. (2021). A deep convolutional neural network for the detection of polyps in colonoscopy images. *Biomedical Signal Processing and Control*, *68*, 102654. https://doi.org/10.1016/j.bspc.2021.102654

Resnik, D. B. (2020). Standards of evidence for Institutional Review Board decision-making. *Accountability in Research, 28*(7), 428-455. https://doi.org/10.1080/08989621.2020.1855149

Rosa, J. P., Guerra, D. J., Horta, N. C., Martins, R. M., & Lourenço, N. C. (2019). Overview of artificial neural networks. *Using Artificial Neural Networks for Analog Integrated Circuit Design Automation*, 21-44. https://doi.org/10.1007/978-3-030-35743-6_3

Russo, A., Gentilini Cacciola, E., Borrazzo, C., Filippi, V., Bucci, T., Vullo, F., Celani, L., Binetti, E., Battistini, L., Ceccarelli, G., Alessandroni, M., Galardo, G., Mastroianni, C. M., & D'Ettorre, G. (2021). Clinical characteristics and outcome of patients with suspected COVID-19 in emergency department (Resiliency study II). *Diagnostics*, *11*(8), 1368. https://doi.org/10.3390/diagnostics11081368

Sai Ganesh, Arisetty & Potnuru, Monika & Tangudu, Naresh. (2023). Artificial Intelligence Tools and Their Usage in Health Care Access and Quality.

Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306. https://doi.org/10.1016/j.physd.2019.132306

Shilo, S., Rossman, H., & Segal, E. (2020). Axes of a revolution: Challenges and promises of big data in healthcare. *Nature Medicine*, *26*(1), 29-38. https://doi.org/10.1038/s41591-019-0727-5

Stark, Z., Lunke, S., Brett, G. R., Tan, N. B., Stapleton, R., Kumble, S., ... & Melbourne Genomics Health Alliance. (2018). Meeting the challenges of implementing rapid genomic testing in acute pediatric care. *Genetics in Medicine*, *20*(12), 1554-1563.

Strijker, D., Bosworth, G., & Bouter, G. (2020). Research methods in rural studies: Qualitative, quantitative, and mixed methods. *Journal of Rural Studies, 78*, 262-270.https://doi.org/10.1016/j.jrurstud.2020.06.007

Sakshihooda, Miss & Mann, Suman. (2021). REVIEW ON PREDICTING DISEASE SEVERITY -LEARNING ALGORITHMS AS CLASSIFIERS FOR DATA WAREHOUSE ENVIRONMENTS.

Sanguineti, M. (2021, July 6). *Plot a TensorFlow Model with Keras Functional API*. Medium. https://towardsdatascience.com/plot-a-tensorflow-model-with-keras-functional-api-f2db639dbbd8#:~:text=We%20will%20then%20plot%20our

Shah, S. (2020). The technological impact of COVID-19 on theFuture of education and health care delivery. *Pain Physician*, *4S;23*(8;4S), S367-S380. https://doi.org/10.36076/ppj.2020/23/s367

Sheikh, A., Anderson, M., Albala, S., Casadei, B., Franklin, B. D., Richards, M., Taylor, D., Tibble, H., & Mossialos, E. (2021). Health information technology and digital innovation for national learning health and care systems. *The Lancet Digital Health*, *3*(6), e383-e396. https://doi.org/10.1016/s2589-7500(21)00005-4

Sinha, H., Awasthi, V., & Ajmera, P. K. (2020). Audio classification using braided convolutional neural networks. *IET Signal Processing*, *14*(7), 448-454.

Sigurdsson, S. (2020). undefined. *Proceedings of the First Workshop on Scholarly Document Processing*. https://doi.org/10.18653/v1/2020.sdp-1.2

Silva, P. B., Andrade, M., & Ferreira, S. (2020). Machine learning applied to road safety modeling: A systematic literature review. *Journal of Traffic and Transportation Engineering (English Edition)*, *7*(6), 775-790. https://doi.org/10.1016/j.jtte.2020.07.004

Suman Mann, M. S. (2021). Review on predicting disease severity – learning algorithms as classifiers for data warehouse environments. *INFORMATION TECHNOLOGY IN INDUSTRY*, *9*(1), 1079-1084. https://doi.org/10.17762/itii.v9i1.239

Sezgin, E., Hussain, S., Rust, S., & Huang, Y. (2023). Extracting medical information from free-text and unstructured patient-generated health data using natural language

processing methods: Feasibility study with real-world data. *JMIR Formative Research*, *7*, e43014. https://doi.org/10.2196/43014

Shehata, M., & Elhosseini, M. (2024). Charting new frontiers: Insights and future directions in ML and DL for image processing. *Electronics*, *13*(7), 1345. https://doi.org/10.3390/electronics13071345

Spasic, I., & Nenadic, G. (2020). undefined. *JMIR Medical Informatics*, *8*(3), e17984. https://doi.org/10.2196/17984

Suman Mann, M. S. (2021). Review on predicting disease severity – learning algorithms as classifiers for data warehouse environments. *INFORMATION TECHNOLOGY IN INDUSTRY*, *9*(1), 1079-1084. https://doi.org/10.17762/itii.v9i1.239

Sadaf, K., Sultana, J., & Ahmad, N. (2023). An xgboost-based classification method to classify breast cancer. *Applications of Machine Learning and Deep Learning on Biological Data*, 75-86. https://doi.org/10.1201/9781003328780-5

Tang, R., Chuang, Y. N., & Hu, X. (2023). The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*. https://arxiv.org/abs/2303.07205

Tahvili, S., Hatvani, L., Ramentol, E., Pimentel, R., Afzal, W., & Herrera, F. (2020). A novel methodology to classify test cases using natural language processing and imbalanced learning. *Engineering Applications of Artificial Intelligence*, *95*, 103878. https://doi.org/10.1016/j.engappai.2020.103878

The Essential Guide to Quality Training Data for Machine Learning. (2024). Cloudfactory.com. https://www.cloudfactory.com/training-data-guide#introduction

Thompson, A. P., Aktulga, H. M., Berger, R., Bolintineanu, D. S., Brown, W. M., Crozier, P. S., In 't Veld, P. J., Kohlmeyer, A., Moore, S. G., Nguyen, T. D., Shan, R., Stevens, M. J., Tranchida, J., Trott, C., & Plimpton, S. J. (2022). LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and

continuum scales. *Computer Physics Communications, 271*, 108171. https://doi.org/10.1016/j.cpc.2021.108171

Timperley, C. S., Herckis, L., Le Goues, C., & Hilton, M. (2021). Understanding and improving artifact sharing in software engineering research. *Empirical Software Engineering*, *26*(4). https://doi.org/10.1007/s10664-021-09973-5

Taye, M. M. (2023). Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. Computers, 12(5), 91–91. https://doi.org/10.3390/computers12050091

u, L., Zhang, J., Xie, Y., Gao, F., Xu, S., Wu, X., & Ye, Z. (2020). Wearable health devices in health care: Narrative systematic review. *JMIR mHealth and uHealth, 8*(11), e18907. https://doi.org/10.2196/18907

Thakkar, A., & Lohiya, R. (2021). A survey on intrusion detection system: Feature selection, model, performance measures, application perspective, challenges, and future research directions. *Artificial Intelligence Review, 55*(1), 453-563. https://doi.org/10.1007/s10462-021-10037-9

Turan, O., Arpınar Yigitbas, B., Turan, P. A., & Mirici, A. (2021). Clinical characteristics and outcomes of hospitalized COVID-19 patients with COPD. *Expert Review of Respiratory Medicine, 15*(8), 1069-1076. https://doi.org/10.1080/17476348.2021.1923484

Universal Rules for Fooling Deep Neural Networks based Text Classification. (n.d.). Ieeexplore.ieee.org. Retrieved October 24, 2023, from https://ieeexplore.ieee.org/abstract/document/8790213

Urcia, I. A. (2021). Comparisons of adaptations in grounded theory and phenomenology: Selecting the specific qualitative research methodology. *International Journal of*

Qualitative Methods, *20*, 160940692110454. https://doi.org/10.1177/16094069211045474

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of k-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-10358-x

Verleye, K. (2019). Designing, writing-up and reviewing case study research: an equifinality perspective. *Journal of Service Management*, *30*(5), 549-576.

Vaci, N., Liu, Q., Kormilitzin, A., De Crescenzo, F., Kurtulmus, A., Harvey, J., O'Dell, B., Innocent, S., Tomlinson, A., Cipriani, A., & Nevado-Holgado, A. (2020). Natural language processing for structuring clinical text data on depression using UK-CRIS. *Evidence Based Mental Health*, *23*(1), 21-26. https://doi.org/10.1136/ebmental-2019-300134

Viani, N., Botelle, R., Kerwin, J., Yin, L., Patel, R., Stewart, R., & Velupillai, S. (2021). undefined. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-020-80457-0

Vijayarani, D.S., & Dhayanand, M.S. (2015). Liver Disease Prediction using SVM and Naïve Bayes Algorithms.

Verbakel, J. Y., Steyerberg, E. W., Uno, H., De Cock, B., Wynants, L., Collins, G. S., & Van Calster, B. (2020). ROC curves for clinical prediction models Part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *Journal of Clinical Epidemiology*, *126*, 207-216. https://doi.org/10.1016/j.jclinepi.2020.01.028

Villalobos-Arias, L., Quesada-López, C., Guevara-Coto, J., Martínez, A., & Jenkins, M. (2020). Evaluating hyper-parameter tuning using random search in support vector machines for software effort estimation. *Proceedings of the 16th ACM International*

Conference on Predictive Models and Data Analytics in Software Engineering. https://doi.org/10.1145/3416508.3417121

Voigtlaender, F. (2023). The universal approximation theorem for complex-valued neural networks. *Applied and Computational Harmonic Analysis, 64*, 33-61. https://doi.org/10.1016/j.acha.2022.12.002

Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., ... & Xu, H. (2020). Deep Learning in clinical natural language processing: a systematic review. *Journal of the American Medical Informatics Association, 27*(3), 457-470.

Wang, D., Su, J., & Yu, H. (2020). Feature extraction and analysis of natural language processing for deep learning English language. *IEEE Access*, *8*, 46335-46345. https://doi.org/10.1109/access.2020.2974101

Wang, J., Huang, J. X., Tu, X., Wang, J., Huang, A. J., Rahman, T., & Bhuiyan, A. (2024). Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. *ACM Computing Surveys*, *56*(7), 1–33. https://doi.org/10.1145/3648471

Wang, H., Li, D., Sheng, T., Sheng, J., Jing, P., & Zhang, D. (2023). A modeling of human reliability analysis on dam failure caused by extreme weather. *Applied Sciences*, *13*(23), 12968. https://doi.org/10.3390/app132312968

Wang, Y., & Wang, J. (2020). Modelling and prediction of global non-communicable diseases. *BMC Public Health, 20*(1). https://doi.org/10.1186/s12889-020-08890-4

What are the advantages and disadvantages of using long short-term memory (LSTM) cells over simple RNN cells? (n.d.). Www.linkedin.com. https://www.linkedin.com/advice/1/what-advantages-disadvantages-using-long-short-term

Wazirali, R. (2020). An improved intrusion detection system based on KNN Hyperparameter tuning and cross-validation. *Arabian Journal for Science and Engineering*, *45*(12), 10859-10873. https://doi.org/10.1007/s13369-020-04907-7

Wu, T., Swaminathan, G., Li, Z., Ravichandran, A., Vasconcelos, N., Bhotika, R., & Soatto, S. (2022). Class-incremental learning with strong pre-trained models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 114*, 9591-9600. https://doi.org/10.1109/cvpr52688.2022.00938

Xie, X., Niu, J., Liu, X., Chen, Z., Tang, S., & Yu, S. (2021). A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, *69*, 101985. https://doi.org/10.1016/j.media.2021.101985

Yamada, A., Niikura, R., Otani, K., Aoki, T., & Koike, K. (2020). Automatic detection of colorectal neoplasia in wireless colon capsule endoscopic images using a deep convolutional neural network. *Endoscopy*, *53*(08), 832-836. https://doi.org/10.1055/a-1266-1066

Yan, C., Yan, Y., Wan, Z., Zhang, Z., Omberg, L., Guinney, J., Mooney, S. D., & Malin, B. A. (2022). A multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications, 13*(1). https://doi.org/10.1038/s41467-022-35295-1

Yoo, S. H., Geng, H., Chiu, T. L., Yu, S. K., Cho, D. C., Heo, J., Choi, M. S., Choi, I. H., Cung Van, C., Nhung, N. V., Min, B. J., & Lee, H. (2020). Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging. *Frontiers in Medicine, 7*. https://doi.org/10.3389/fmed.2020.00427

Yousefi, M., Tabatabaei, S. H., Rikhtehgaran, R., Pour, A. B., & Pradhan, B. (2021). Application of Dirichlet process and support vector machine techniques for mapping

alteration zones associated with porphyry copper deposit using ASTER remote sensing imagery. *Minerals*, *11*(11), 1235. https://doi.org/10.3390/min11111235

Yu, J., Choi, J. S., Giannoni, C., Patel, A. J., & Gallagher, K. K. (2019). Juvenile Nasopharyngeal Angiofibroma outcomes and cost: Analysis of the kids' inpatient database. *Annals of Otology, Rhinology & Laryngology*, *129*(5), 498-504. https://doi.org/10.1177/0003489419896597

Yu, N., Xu, K., Chen, K., Liu, S., Zheng, T., & Song, M. (2024). Multi-channel graph fusion representation for tabular data imputation. *2024 International Joint Conference on Neural Networks (IJCNN)*, *33*, 1-8. https://doi.org/10.1109/ijcnn60899.2024.10651425

Zarei, E., Khan, F., & Abbassi, R. (2022). A dynamic human-factor risk model to analyze safety in sociotechnical systems. *Process Safety and Environmental Protection*, *164*, 479-498. https://doi.org/10.1016/j.psep.2022.06.040

Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: A narrative review. *npj Digital Medicine*, *5*(1). https://doi.org/10.1038/s41746-022-00589-7

Zhang, L., Wen, J., Li, Y., Chen, J., Ye, Y., Fu, Y., & Livingood, W. (2021). A review of machine learning in building load prediction. *Applied Energy*, *285*, 116452. https://doi.org/10.1016/j.apenergy.2021.116452

Zhong, Y., Chalise, P., & He, J. (2020). Nested cross-validation with ensemble feature selection and classification model for high-dimensional biological data. *Communications in Statistics - Simulation and Computation*, *52*(1), 110-125. https://doi.org/10.1080/03610918.2020.1850790

Appendix A (Source Code)

audio_medical_diagnosis.pdf

text_medical_diagnosis.pdf

multimodal_medical_diagnosis.pdf